

Accuracy guaranties for ℓ_1 recovery of block-sparse signals

Anatoli Juditsky* Fatma Kılınç Karzan† Arkadi Nemirovski‡ Boris Polyak§

November 11, 2011

Abstract

We discuss new methods for the recovery of signals with block-sparse structure, based on ℓ_1 -minimization. Our emphasis is on verifiable conditions on the problem parameters (sensing matrix and the block structure) for accurate recovery and efficiently computable bounds for the recovery error. These bounds are then optimized with respect to the method parameters to construct the estimators with improved statistical properties. To justify the proposed approach we provide an oracle inequality which links the properties of the recovery algorithms and the best estimation performance. We also propose a new matching pursuit algorithm for block-sparse recovery.

1 Introduction

The problem we consider in this paper is to estimate a linear transform $Bx \in \mathbb{R}^N$ of a vector $x \in \mathbb{R}^n$ from the observations

$$y = Ax + u + \xi. \quad (1.1)$$

Here A is a given $m \times n$ sensing matrix, B is a given $N \times n$ matrix, and $u + \xi$ is the observation error; in this error, u is an unknown *nuisance* known to belong to a given compact convex set $\mathcal{U} \subset \mathbb{R}^m$ symmetric w.r.t. the origin, and ξ is random noise with known distribution P .

We assume that the space \mathbb{R}^N where Bx lives is represented as $\mathbb{R}^N = \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_K}$, so that a vector $w \in \mathbb{R}^N$ is a block vector: $w = [w[1]; \dots; w[K]]$ with blocks $w[k] \in \mathbb{R}^{n_k}$, $1 \leq k \leq K$. In particular, $Bx = [B[1]x; \dots; B[K]x]$ with $n_k \times n$ matrices $B[k]$, $1 \leq k \leq K$. While we do not assume that the vector x is sparse in the usual sense, we do assume that the linear transform Bx to be estimated is *s-block sparse*, meaning that at most a given number s of the blocks $B[k]x$, $1 \leq k \leq K$, are nonzero.

The recovery routines we intend to consider are based on *block- ℓ_1 minimization*, i.e., the estimate $\hat{w}(y)$ of $w = Bx$ is $B\hat{z}(y)$, where $\hat{z}(y)$ is obtained by minimizing the norm $\sum_{k=1}^K \|B[k]z\|_{(k)}$ over signals $z \in \mathbb{R}^n$ with Az “fitting,” in certain precise sense, the observations y . Above, $\|\cdot\|_{(k)}$ are given in advance norms on the spaces \mathbb{R}^{n_k} where the blocks of Bx take their values.

In the sequel we refer to the given in advance collection $(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ as to the *representation structure* (r.s.). Given such a structure and A , our ultimate goal is to understand how well can we recover the *s-block-sparse* transform Bx by appropriately implemented block ℓ_1 minimization.

*LJK, Université J. Fourier, B.P. 53, 38041 Grenoble Cedex 9, France, Anatoli.Juditsky@imag.fr

†Georgia Institute of Technology, Atlanta, Georgia 30332, USA, fkilinc@isye.gatech.edu

‡Georgia Institute of Technology, Atlanta, Georgia 30332, USA, nemirovs@isye.gatech.edu

Research of the second and the third authors was supported by the Office of Naval Research grant #240667V.

§Institute of Control Sciences of Russian Academy of Sciences, Moscow 117997, Russia, boris_ipu.rssi.ru

Related Compressed Sensing research Our situation and goal form a straightforward extension of the usual sparsity/block sparsity Compressed Sensing framework. Indeed, the *standard representation structure* with $B = I_n$, $n_k = 1$, and $\|\cdot\|_{(k)} = |\cdot|$, $1 \leq k \leq K = n$, leads to the standard Compressed Sensing setting – recovering a sparse signal $x \in \mathbb{R}^n$ from its noisy observations (1.1) via ℓ_1 minimization. The case of nontrivial block structure $\{n_k, \|\cdot\|_{(k)}\}_{k=1}^K$ and $B = I$ is generally referred to as *block-sparse*, and has been considered in numerous recent papers. Specifically, there seem to be a number of applications where block-sparsity (with $B = I_n$) arises naturally (see, e.g., [16] and references therein), such as multi-band signals, measurements of gene expression levels, or estimation of multiple measurement vectors sharing a joint sparsity pattern, among many others. Several methods of estimation and selection extending the “plain” ℓ_1 -minimization to block sparsity were proposed and investigated recently. Most of the related research focused so far on *block regularization schemes* — group Lasso recovery

$$\hat{x}(y) \in \underset{z=[z^1; \dots; z^K] \in \mathbb{R}^n = \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_K}}{\text{Argmin}} \left\{ \|Az - y\|_2^2 + \lambda \sum_{k=1}^K \|z[k]\|_2 \right\}$$

(here $\|\cdot\|_2$ is the Euclidean norm of the block). In particular, the literature on “plain Lasso” (the case of $n_k = 1$, $1 \leq k \leq K = n$) has a important counterpart on group Lasso, see, e.g., [2, 4, 12, 14, 15, 16, 18, 19, 24, 27, 28, 29, 30, 31, 33], and references therein. Another celebrated technique of sparse recovery, Dantzig selector, originating from [9], has also received its counterpart for recovery of block-sparse signals, which is dealt with in [20, 25]. Most of the cited papers focus on bounding recovery errors in terms of magnitude of the observation noise and “ s -concentration” of the true signal x (the distance from the space of signals with at most s nonzero blocks — the sum of magnitudes $\|x[k]\|_2$ of all but the s largest in magnitude blocks in x). Typically, these results rely on natural block analogy (“Block RIP,” see, e.g., [16]) of the celebrated Restricted Isometry Property introduced by Candés and Tao [11, 10], or on block analogies [26] of the Restricted Eigenvalue Property introduced in [6].

Contributions of this paper The first (by itself, minor) novelty in our problem setting is the presence of the linear mapping B . We are not aware of any preceding work handling the case of a “nontrivial” (i.e., different from the identity) B . We qualify this novelty as minor, since in fact the case of a nontrivial B can be reduced to the one of $B = I$ ¹. However, “can be reduced” is not the same as “should be reduced,” since nontrivial B ’s arise naturally in many applications. This is the case, e.g., when x is the solution of a linear finite-difference equation with sparse right hand side (“evolution of a linear plant corrected from time to time by impulse control”), where B is the matrix of the corresponding finite-difference operator. We believe that introducing B adds some useful flexibility (and as a matter of fact costs nothing, as far as the theoretical analysis is concerned).

We believe, however, that the major novelty in what follows is the emphasis on *verifiable* conditions on A and the r.s. which *guarantee* good recovery of the transform Bx from noisy observations of Ax , provided that the transform in question is nearly s -block sparse, and the observation noise is low. Note that such guarantees cannot be obtained from the “classical” conditions used when studying theoretical properties of block-sparse recovery (with a notable exception of the Mutual Block-Incoherence condition of [15]). The latter means that given the matrix A , one cannot answer in any reasonable time if the (Block-) Restricted Isometry or Restricted Eigenvalue property hold with given parameters. While the efficient verifiability is by no means necessary for a condition to be meaningful and useful, we believe also that verifiability has its value and is worthy of being investigated. In particular, the verifiability allows to design new recovery routines with explicit confidence bounds for the recovery error and then optimize these bounds with respect to the parameters of the recovery. In this respect, the current work extends the results of [23, 21, 22], where

¹Assuming, e.g., that $x \mapsto Bx$ is an “onto” mapping, we can treat Bx as our signal, the observations being Py , where P is the projector onto the orthogonal complement to the linear subspace $A \cdot \text{Ker} B$ in \mathbb{R}^m ; with $y = Ax + u + \xi$, we have $Py = GBx + P(u + \xi)$ with an explicitly given matrix G .

ℓ_1 -recovery of the “usual” sparse vectors was considered (in the first two papers – in the case of uncertain-but-bounded observation errors, and in the third – in the case of Gaussian observation noise). Precisely, we propose here new routines of block-sparse recovery which explicitly utilize the *verifiability certificate* – the *contrast matrix*, and show how these routines may be tuned to attain the best performance bounds.

To give an impression of what will follow, we present here a short summary of our major results. To streamline this summary, we restrict ourselves for the time being with the case where (a) the random noise ξ in (1.1) is Gaussian: $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ with known $\sigma^2 > 0$, and (b) all the norms $\|\cdot\|_{(k)}$ are just $\|\cdot\|_r$ -norms, with the value of r common for all $1 \leq k \leq K$. Let s be a given positive integer — an a priori upper bound on the number of nonzero blocks $B[k]x$ in the transforms we intend to recover well, and $\epsilon \ll 1$ be the a given tolerance. We fix an $m \times n$ sensing matrix A and an r.s. $(B, n_1, \dots, n_K, \|\cdot\|_r, \dots, \|\cdot\|_r)$.

Condition $\mathbf{Q}_{s,q}$ Given s and $q \in [1, \infty]$, we introduce a condition $\mathbf{Q}_{s,q}$ on an $m \times N$ contrast matrix H , specifically, the condition

$$\forall(x \in \mathbb{R}^n) : L_{s,q}(Bx) \leq s^{\frac{1}{q}} L_\infty(H^T Ax) + \frac{1}{4} s^{\frac{1}{q}-1} L_1(Bx)$$

where for $w = [w[1]; \dots; w[K]] \in \mathbb{R}^N$ and $p \in [1, \infty]$, $L_p(w) = \|[\|w[1]\|_r; \dots; \|w[K]\|_r]\|_p$ is the norm of w ; $L_{s,p}(w)$ is the norm of w obtained as follows: we zero out all but the s largest in “magnitude” $\|w[k]\|_r$ blocks in w , and take the L_p -norm of the resulting s -block-sparse vector. For example, $L_{s,\infty}(w)$ is, independently of s , the maximum of magnitudes $\|w[k]\|_r$ of blocks in w .

Recovery routines Given an $\epsilon > 0$ and an $m \times N$ contrast matrix $H = [h^1, \dots, h^N]$, we introduce two recovery routines: *regular L_1 recovery* (cf. (block-) Dantzig selector)

$$\begin{aligned} \hat{x}_{\text{reg}}(y) &\in \underset{z \in \mathbb{R}^n}{\text{Argmin}} \left\{ L_1(Bz) : \|H^T(y - Az)\|_\infty \leq \nu(H) \right\}, \\ \nu(H) &= \max_{1 \leq j \leq N} \left[\max_{u \in \mathcal{U}} u^T h^j + \sigma \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|h^j\|_2 \right], \end{aligned} \quad (1.2)$$

$\text{Erfinv}(\cdot)$ being the inverse error function², and *penalized L_1 recovery* (cf. (block-) Lasso)

$$\hat{x}_{\text{pen}}(y) \in \underset{z \in \mathbb{R}^n}{\text{Argmin}} \left[L_1(Bz) + 2s \|H^T(y - Az)\|_\infty \right].$$

Note that the regular L_1 recovery can be undefined; this happens when the corresponding optimization problem is infeasible. The penalized recovery always is well defined.

Error bounds for regular and penalized recoveries Our main related result is as follows (see Theorems 3.1, 3.3): *Let a contrast matrix H satisfy the condition $\mathbf{Q}_{s,q}$. Then there exists a set Ξ of realizations of ξ such that $\text{Prob}\{\xi \in \Xi\} \geq 1 - \epsilon$ and for all $\xi \in \Xi$, $x \in \mathbb{R}^n$ and $u \in \mathcal{U}$, $\hat{x}_{\text{reg}}(Ax + u + \xi)$ is well defined, and for both $\hat{x} = \hat{x}_{\text{reg}}(Ax + u + \xi)$ and $\hat{x} = \hat{x}_{\text{pen}}(Ax + u + \xi)$ one has*

$$\forall p \in [1, q] : L_p(B\hat{x} - Bx) \leq 4(2s)^{\frac{1}{p}} [\nu(H) + s^{-1} v_s(Bx)] \quad (1.3)$$

where $v_s(w)$ is the “ s -concentration of w ,” that is, the sum of magnitudes $\|w[k]\|_r$ of all but the s largest in magnitude blocks in w . Note that for the case of the standard r.s., the corresponding constructions and results were developed in [22].

²i.e., $u = \text{Erfinv}(\delta)$ means that $\frac{1}{\sqrt{2\pi}} \int_u^\infty e^{-t^2/2} dt = \delta$.

Verifiable sufficient condition for $\mathbf{Q}_{s,q}$ and contrast optimization Similarly to the plain and block Restricted Isometry/Eigenvalue Properties, condition $\mathbf{Q}_{s,q}$ is computationally intractable. In other words, given a candidate contrast matrix H , it is difficult to verify whether it satisfies or does not satisfy $\mathbf{Q}_{s,q}$. We, however, can point out a *verifiable sufficient condition* for H to satisfy $\mathbf{Q}_{s,q}$. Specifically, we show (Proposition 5.1) that *H definitely satisfies $\mathbf{Q}_{s,q}$, if there exists a $N \times N$ matrix V (which we treat as a $K \times K$ block matrix with $n_k \times n_\ell$ blocks $V^{k\ell}$) such that*

$$(a): \quad B = VB + H^T A, \text{ and } (b): \quad \|[\|V^{1\ell}\|_{r,r}; \|V^{2\ell}\|_{r,r}; \dots; \|V^{K\ell}\|_{r,r}]\|_{s,q} \leq \frac{1}{4}s^{\frac{1}{q}-1}, \quad (1.4)$$

where $\|V^{k\ell}\|_{r,r} = \max_{u \in \mathbb{R}^{n_\ell}} \{\|V^{k\ell}u\|_r : \|u\|_r \leq 1\}$, and $\|u\|_{s,p}$ is the norm on \mathbb{R}^K defined as follows: we zero out all but the s largest in magnitude entries in vector u , and take the $\|\cdot\|_p$ -norm of the resulting vector.

One can use the above sufficient condition in order to build a “quasi-optimal” contrast matrix, specifically, by minimizing $\nu(H)$, defined in (1.2), over pairs (V, H) satisfying the system of convex constraints (1.4) (provided, of course, that this system of constraints is feasible). The resulting problem is computationally tractable, provided that the matrix norms $\|\cdot\|_{r,r}$ are efficiently computable, which indeed is the case when $r = 1$, or $r = 2$, or $r = \infty$.

Verifiable sufficient condition in the case $q = \infty$ In general, the proposed verifiable (at least for $r \in \{1, 2, \infty\}$) sufficient condition for H to satisfy $\mathbf{Q}_{s,q}$ is *not* necessary, and the condition $\mathbf{Q}_{s,q}$ itself seems to be intractable. There exists, however, a notable exception – this is the case of $q = \infty$ and $r = \infty$. We show (Proposition 4.1) that here the verifiable sufficient condition is *necessary and sufficient* for H to satisfy $\mathbf{Q}_{s,\infty}$. Moreover, the latter condition is “fully computationally tractable,” meaning that one can optimize efficiently the quantity $\nu(H)$ over the contrast matrices H satisfying $\mathbf{Q}_{s,\infty}$, thus ending up with an optimal, as far as the error bound (1.3) is concerned, recovery routines. Note that when $q = \infty$, the bound (1.3) holds true in the largest possible range $1 \leq p \leq \infty$ of values of p .

In the case of the standard r.s., the sufficient condition (1.4) reduces to the verifiable sufficient condition for the validity of ℓ_1 recovery established in [23]. As we have mentioned above, the only known so far *verifiable* sufficient condition for the validity of *block* ℓ_1 recovery of block-sparse signals is the Mutual Block-Incoherence condition (cf. [15] and [17]) dealing with the case of $B = I$ and $r = 2$. This condition is a block analogy of the usual mutual incoherence condition originating from [13]. We show in Section 5.4 that the Mutual Block-Incoherence condition is “covered” by the case of $B = I$, $r = 2$ of the verifiable condition (1.4).

Oracle inequality in the case $q = \infty$, $r = \infty$ As the majority of good error bounds in Compressed Sensing, the error bound (1.3) expresses the following quite intuitive fact. Imagine that instead of indirect observation (1.1) of a transform $w = Bx$, we were observing this transform directly with noise: $y = w + \zeta$. Here the observation error ζ is such that with probability $\geq 1 - \epsilon$ one has $L_\infty(\zeta) \leq \nu$. It is easily seen that in the latter case, *in the range $v_s(w) \leq s\nu(H)$ of s -concentrations of w , the best $(1 - \epsilon)$ -reliable bound on the $L_p(\cdot)$ -norm of the recovery error of w coincides, within an absolute constant factor, with the right hand side of (1.3).* Thus, a natural interpretation of the error bound (1.3) is that *as far as recovery of transforms Bx with s -concentration $v_s(Bx) \leq s\nu(H)$ is concerned, everything is as if we were given a direct observation of Bx contaminated with a noise of typical L_∞ -magnitude $\leq \nu(H)$.* One of the main results of this paper is that, to some extent, the opposite also is true, provided that $r = \infty$ and (1.3) holds true in the entire range $1 \leq p \leq \infty$ of values of p . Specifically, we prove (see Proposition 4.2) the following. *Let all the block norms be the $\|\cdot\|_\infty$ -norms, and let the observation error be present (that is, either $\sigma > 0$, or \mathcal{U} contains a neighborhood of the origin). Let, further, for some integer S and positive ν there exist a routine (an oracle) $\hat{w}(y) \equiv \hat{B}x(y)$ for recovering Bx from observations (1.1) such that*

$$\forall(u \in \mathcal{U}, x \in \mathbb{R}^n : v_S(Bx) \leq S\nu) : \text{Prob}_{\xi \sim \mathcal{N}(0, I)} \{L_\infty(B[x - \hat{x}(Ax + u + \sigma\xi)]) \leq \nu + S^{-1}v_S(Bx)\} \geq 1 - \epsilon.$$

(cf. (1.3) with $p = \infty$). Then for every integer s , $1 \leq s \leq \frac{S}{8}$, there exists a contrast matrix $H \in \mathbb{R}^{m \times N}$ and a “certificate” $V = [V^{k\ell}]_{k,\ell=1}^K \in \mathbb{R}^{N \times N}$ such that

$$B = VB + H^T A, \quad \|V^{k\ell}\|_{\infty, \infty} \leq \frac{1}{4s}, \quad 1 \leq k, \ell \leq K, \quad \text{and} \quad \nu(H) \leq \nu^* := 2\nu \frac{\text{Erfinv}(\frac{\epsilon}{2N})}{\text{Erfinv}(\frac{\epsilon}{2})}.$$

In other words, when ϵ is small, the condition (1.4) is satisfied by an appropriate H for all s in the range $[1, s^*]$, such that s^* and $\nu(H)$ coincide, within some absolute constant factors, with S and ν , respectively.

All proofs are placed in the Appendix.

2 Problem statement

Notation. In the sequel, we deal with

- *signals* – vectors $x = [x_1; \dots; x_n] \in \mathbb{R}^n$, and a $m \times n$ sensing matrix A ;
- *representations of signals* – block vectors $w = [w[1]; \dots; w[K]] \in \mathcal{W} := \mathbb{R}_{w[1]}^{n_1} \times \dots \times \mathbb{R}_{w[K]}^{n_K}$, and the representation matrix $B = [B[1]; \dots; B[K]]$, $B[k] \in \mathbb{R}^{n_k \times n}$; the representation of a signal $x \in \mathbb{R}^n$ is the block vector $w = Bx$ with the blocks $B[1]x, \dots, B[K]x$.

From now on, the dimension of \mathcal{W} is denoted by N :

$$N = n_1 + \dots + n_K.$$

The factors \mathbb{R}^{n_k} of the representation space \mathcal{W} are equipped with norms $\|\cdot\|_{(k)}$; the conjugate norms are denoted by $\|\cdot\|_{(k,*)}$. A vector $w = [w[1]; \dots; w[K]]$ from \mathcal{W} is called *s-block-sparse*, if the number of nonzero blocks $w[k] \in \mathbb{R}^{n_k}$ in w is at most s . A vector $x \in \mathbb{R}^n$ will be called *s-block-sparse*, if its representation Bx is so. We refer to the collection $(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ as the *representation structure* (r.s. for short).

For $w \in \mathcal{W}$, we call the number $\|w[k]\|_{(k)}$ the *magnitude* of the k -th block in w , and denote by w^s the representation vector obtained from w by zeroing out all but the s largest in magnitude blocks in w (with the ties resolved arbitrarily). For $I \subset \{1, \dots, K\}$ and a representation vector w , w_I denotes the vector obtained from w by keeping intact the blocks $w[k]$ with $k \in I$ and zeroing out all remaining blocks. For $w \in \mathcal{W}$ and $1 \leq p \leq \infty$, we denote by $L_p(w)$ the $\|\cdot\|_p$ -norm of the vector $[\|w[1]\|_{(1)}; \dots; \|w[K]\|_{(K)}]$, so that $L_p(\cdot)$ is a norm on \mathcal{W} with the conjugate norm $L_p^*(w) = \|[\|w[1]\|_{(1,*)}; \dots; \|w[K]\|_{(K,*)}]\|_{p^*}$, $p^* = \frac{p}{p-1}$. Given positive integer $s \leq K$, we set $L_{s,p}(w) = L_p(w^s)$. Note that $L_{s,p}(\cdot)$ is a norm on \mathcal{W} .

Problem of interest is as follows: given an observation

$$y = Ax + u + \xi, \tag{2.5}$$

of unknown signal $x \in \mathbb{R}^n$, we want to recover the representation Bx of x , knowing in advance that this representation is “nearly s -block-sparse,” that is, the representation can be approximated by an s -block-sparse one; the L_1 -error of this approximation will enter our error bounds.

In (2.5), the term $u + \xi$ is the observation error; in this error, u is an unknown *nuisance* known to belong to a given compact convex set $\mathcal{U} \subset \mathbb{R}^m$ symmetric w.r.t. the origin, and ξ is random noise with known distribution P .

Condition $\mathbf{Q}_{s,q}(\kappa)$ We start with introducing the condition which will be instrumental in all subsequent constructions and results. Let a sensing matrix A and an r.s. $\mathcal{S} = (B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ be given, and let $s \leq K$ be a positive integer, $q \in [1, \infty]$ and $\kappa \geq 0$. We say that a pair $(H, \|\cdot\|)$, where $H \in \mathbb{R}^{m \times M}$ and $\|\cdot\|$ is a norm on \mathbb{R}^M , satisfies the condition $\mathbf{Q}_{s,q}(\kappa)$ associated with the matrices A, B and the r.s., if

$$\forall x \in \mathbb{R}^n : L_{s,q}(Bx) \leq s^{\frac{1}{q}} \|H^T Ax\| + \kappa s^{\frac{1}{q}-1} L_1(Bx). \quad (2.6)$$

The following observation is evident:

Observation 2.1 *Given A and an r.s. $(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$, let $(H, \|\cdot\|)$ satisfy $\mathbf{Q}_{s,q}(\kappa)$. Then $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q'}(\kappa')$ for all $q' \in (1, q)$ and $\kappa' \geq \kappa$. Besides this, if $s' \leq s$ is a positive integer, $((s/s')^{\frac{1}{q}} H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s',q}((s'/s)^{1-\frac{1}{q}} \kappa)$. Further, if $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q}(\kappa)$, $q' \geq q$, and κ' and a positive integer s' are such that $\kappa'(s')^{\frac{1}{q'}-1} \geq \kappa s^{\frac{1}{q}-1}$, then $(s^{\frac{1}{q}}(s')^{-\frac{1}{q'}} H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s',q'}(\kappa')$. In particular, when $s' \leq s^{1-\frac{1}{q}}$, the fact that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q}(\kappa)$ implies that $(s^{\frac{1}{q}} H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s',\infty}$.*

Relation to known conditions for the validity of sparse ℓ_1 recovery. Note that whenever

$$(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$$

is the *standard* r.s., meaning that B is the identity matrix, $n_1 = \dots = n_K = 1$ and $\|\cdot\|_{(k)} = |\cdot|$ for all k , the condition $\mathbf{Q}_{s,q}(\kappa)$ reduces to the condition $\mathbf{H}_{s,q}(\kappa)$ introduced in [22]. On the other hand, condition $\mathbf{Q}_{s,p}(\kappa)$ is closely related to known conditions, introduced to study the properties of recovery routines in the context of block-sparsity. Specifically, consider an r.s. with $B = I_n$, and let us make the following observation:

Let $(H, \|\cdot\|_{\infty})$ satisfy $\mathbf{Q}_{s,q}(\kappa)$ and let $\hat{\lambda}$ be the maximum of the Euclidean norms of columns in H . Then

$$\forall x \in \mathbb{R}^n : L_{s,q}(x) \leq \hat{\lambda} s^{\frac{1}{q}} \|Ax\|_2 + \kappa s^{\frac{1}{q}-1} L_1(x). \quad (2.7)$$

Let us fix the r.s. $\mathcal{S}_2 = (I_n, n_1, \dots, n_K, \|\cdot\|_2, \dots, \|\cdot\|_2)$. Condition (2.7) with $\kappa < 1/2$ plays crucial role in the performance analysis of group-Lasso and Dantzig Selector. For example, the error bounds for Lasso recovery obtained in [26] rely upon the Restricted Eigenvalue assumption $\text{RE}(s, \kappa)$ as follows: there is $\kappa > 0$ such that

$$L_2(x^s) \leq \frac{1}{\kappa} \|Ax\|_2 \text{ whenever } 3L_1(x^s) \geq L_1(x - x^s).$$

In this case $L_{s,1}(x) \leq \sqrt{s} L_{s,2}(x) \leq \frac{\sqrt{s}}{\kappa} \|Ax\|_2$ whenever $4L_{s,1}(x) \geq L_1(x)$, so that

$$\forall x \in \mathbb{R}^n : L_{s,1}(x) \leq \frac{s^{1/2}}{\kappa} \|Ax\|_2 + \frac{1}{4} L_1(x) \quad (2.8)$$

what is (2.7) with $q = 1$, $\kappa = 1/4$ and $\hat{\lambda} = (\kappa\sqrt{s})^{-1}$ (observe that (2.8) is nothing but the “block version” of the Compatibility condition from [7]).

Recall that a sensing matrix $A \in \mathbb{R}^{m \times n}$ satisfies the Block Restricted Isometry Property $\text{BRIP}(\delta, k)$ (see, e.g. [16]) with $\delta \geq 0$ and a positive k if for every $x \in \mathbb{R}^n$ with at most k non-vanishing blocks one has

$$(1 - \delta) \|x\|_2^2 \leq x^T A^T A x \leq (1 + \delta) \|x\|_2^2. \quad (2.9)$$

Proposition 2.1 *Let $A \in \mathbb{R}^{m \times n}$ satisfy $\text{BRIP}(\delta, 2s)$ for some $\delta < 1$ and positive integer s . Then*

- (i) *The pair $(H = \frac{s^{-1/2}}{\sqrt{1-\delta}} I_m, \|\cdot\|_2)$ satisfies the condition $\mathbf{Q}_{s,2}(\frac{\delta}{1-\delta})$ associated with A and the r.s. \mathcal{S}_2 .*
- (ii) *The pair $(H = \frac{1}{1-\delta} A, L_{\infty}(\cdot))$ satisfies the condition $\mathbf{Q}_{s,2}(\frac{\delta}{1-\delta})$ associated with A and the r.s. \mathcal{S}_2 .*

Our last observation here is as follows: let $(H, \|\cdot\|)$ satisfy $\mathbf{Q}_{s,q}(\kappa)$, the r.s. being $(B, n_1, \dots, n_K, \|\cdot\|_2, \dots, \|\cdot\|_2)$, and let $d = \max_k n_k$. Then $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q}(\sqrt{d}\kappa)$, the r.s. being $(B, n_1, \dots, n_K, \|\cdot\|_{\infty}, \dots, \|\cdot\|_{\infty})$.

3 Accuracy bounds for ℓ_1 block recovery routines

Throughout this section we fix an r.s. $\mathcal{S} = (B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ and a sensing matrix A .

3.1 Regular ℓ_1 recovery

We define the *regular ℓ_1 recovery* as

$$\hat{x}_{\text{reg}}(y) \in \underset{u}{\text{Argmin}} \{L_1(Bu) : \|H^T(Au - y)\| \leq \rho\}, \quad (3.10)$$

where the *contrast matrix* $H \in \mathbb{R}^{m \times M}$, the norm $\|\cdot\|$ and $\rho > 0$ are parameters of the construction.

Theorem 3.1 *Let s be a positive integer, $q \in [1, \infty]$, $\kappa \in (0, 1/2)$, and $\epsilon \in (0, 1)$. Assume that the pair $(H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_{s,q}(\kappa)$ associated with A and r.s. \mathcal{S} and that there exists a set Ξ satisfying $P(\Xi) \geq 1 - \epsilon$ and*

$$\|H^T(u + \xi)\| \leq \rho \quad \forall (u \in \mathcal{U}, \xi \in \Xi) \quad (3.11)$$

Then for all $x \in \mathbb{R}^n$, $u \in \mathcal{U}$ and $\xi \in \Xi$ one has

$$L_p(B[\hat{x}_{\text{reg}}(Ax + u + \xi) - x]) \leq \frac{4(2s)^{\frac{1}{p}}}{1 - 2\kappa} \left[\rho + \frac{1}{2s} L_1(Bx - [Bx]^s) \right], \quad 1 \leq p \leq q. \quad (3.12)$$

The above result can be slightly strengthened by replacing the assumption that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q}(\kappa)$, $\kappa < 1/2$, with a weaker, by Observation 2.1, assumption that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,1}(\varkappa)$ with $\varkappa < 1/2$ and satisfies $\mathbf{Q}_{s,q}(\kappa)$ with some (perhaps large) κ :

Theorem 3.2 *Given A , r.s. \mathcal{S} , integer $s > 0$, $q \in [1, \infty]$ and $\epsilon \in (0, 1)$, assume that $(H, \|\cdot\|)$ satisfies the condition $\mathbf{Q}_{s,1}(\varkappa)$ with $\varkappa < 1/2$ and the condition $\mathbf{Q}_{s,q}(\kappa)$ with some $\kappa \geq \varkappa$, and let ρ be such that there exists a set Ξ satisfying $P(\Xi) \geq 1 - \epsilon$ and*

$$\|H^T(u + \xi)\| \leq \rho \quad \forall (u \in \mathcal{U}, \xi \in \Xi).$$

Then for all $x \in \mathbb{R}^n$, $u \in \mathcal{U}$, $\xi \in \Xi$ and p , $1 \leq p \leq q$, it holds:

$$L_p(B[\hat{x}_{\text{reg}}(Ax + u + \xi) - x]) \leq \frac{4(2s)^{\frac{1}{p}} [1 + \kappa - \varkappa]^{\frac{q(p-1)}{p(q-1)}}}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1(Bx - [Bx]^s) \right]. \quad (3.13)$$

3.2 Penalized ℓ_1 recovery

Penalized ℓ_1 recovery is

$$\hat{x}_{\text{pen}}(y) \in \underset{u}{\text{Argmin}} \{L_1(Bu) + \lambda \|H^T(Ax - y)\|\}, \quad (3.14)$$

where $H \in \mathbb{R}^{m \times M}$, $\|\cdot\|$ and a positive real λ are parameters of the construction.

Theorem 3.3 *Given A , r.s. \mathcal{S} , integer s , $q \in [1, \infty]$ and $\epsilon \in (0, 1)$, assume that $(H, \|\cdot\|)$ satisfies the conditions $\mathbf{Q}_{s,q}(\kappa)$ and $\mathbf{Q}_{s,1}(\varkappa)$ with $\varkappa < 1/2$ and $\kappa \geq \varkappa$.*

(i) *Let $\lambda \geq 2s$. Then for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ it holds for $1 \leq p \leq q$:*

$$L_p(B[\hat{x}_{\text{pen}}(y) - x]) \leq \frac{4\lambda^{\frac{1}{p}}}{1 - 2\varkappa} \left[1 + \frac{\kappa\lambda}{2s} - \varkappa \right]^{\frac{q(p-1)}{p(q-1)}} \left[\|H^T(Ax - y)\| + \frac{1}{2s} L_1(Bx - [Bx]^s) \right]. \quad (3.15)$$

In particular, with $\lambda = 2s$ we have

$$L_p(B[\hat{x}_{\text{pen}}(y) - x]) \leq \frac{4(2s)^{\frac{1}{p}}}{1 - 2\varkappa} [1 + \kappa - \varkappa]^{\frac{q(p-1)}{p(q-1)}} \left[\|H^T(Ax - y)\| + \frac{1}{2s} L_1(Bx - [Bx]^s) \right], \quad 1 \leq p \leq q. \quad (3.16)$$

(ii) Let $\rho \geq 0$ be such that the set $\Xi = \{\xi : \|H^T(\xi + u)\| \leq \rho \forall u \in \mathcal{U}\}$ satisfies $\text{Prob}\{\xi \in \Xi\} \geq 1 - \epsilon$. Then for all $x \in \mathbb{R}^n$, $u \in \mathcal{U}$ and all $\xi \in \Xi$ one has for $1 \leq p \leq q$:

$$\begin{aligned} \lambda \geq 2s &\Rightarrow L_p(B[\widehat{x}_{\text{pen}}(Ax + u + \xi) - x]) \leq \frac{4\lambda^{\frac{1}{p}}}{1-2\kappa} \left[1 + \frac{\kappa\lambda}{2s} - \kappa\right]^{\frac{q(p-1)}{p(q-1)}} \left[\rho + \frac{1}{2s}L_1(Bx - [Bx]^s)\right], \\ \lambda = 2s &\Rightarrow L_p(B[\widehat{x}_{\text{pen}}(Ax + u + \xi) - x]) \leq \frac{4(2s)^{\frac{1}{p}}}{1-2\kappa} \left[1 + \kappa - \kappa\right]^{\frac{q(p-1)}{p(q-1)}} \left[\rho + \frac{1}{2s}L_1(Bx - [Bx]^s)\right]. \end{aligned} \quad (3.17)$$

Discussion. Let us compare the error bounds of the regular and the penalized ℓ_1 recoveries associated with the same pair $(H, \|\cdot\|)$ satisfying the condition $\mathbf{Q}_{s,q}(\kappa)$ with $\kappa = 1/2$. Let

$$\rho_\epsilon[H, \|\cdot\|] = \min \left\{ \rho : \text{Prob} \left\{ \xi : \|H^T(u + \xi)\| \leq \rho \forall u \in \mathcal{U} \right\} \geq 1 - \epsilon \right\}; \quad (3.18)$$

this is nothing but the smallest ρ meeting the condition (3.11) with Ξ satisfying $\text{Prob}\{\xi \in \Xi\} \geq 1 - \epsilon$ and thus – the smallest ρ for which the error bound (3.12) for the regular ℓ_1 recovery holds true with probability $1 - \epsilon$ (or at least the smallest ρ for which the latter claim is supported by Theorem 3.1). With $\rho = \rho_\epsilon[H, \|\cdot\|]$, the regular ℓ_1 recovery guarantees (and that is the best guarantee one can extract from Theorem 3.1) that

(!) For some set Ξ , $\text{Prob}\{\xi \in \Xi\} \geq 1 - \epsilon$, of “good” realizations of the random component ξ of the observation error, one has

$$L_p(B[\widehat{x}(Ax + u + \xi) - x]) \leq \frac{4(2s)^{\frac{1}{p}}}{1-2\kappa} \left[\rho_\epsilon[H, \|\cdot\|] + \frac{1}{2s}L_1(Bx - [Bx]^s) \right], \quad 1 \leq p \leq q, \quad (3.19)$$

whenever $x \in \mathbb{R}^n$, $u \in \mathcal{U}$, $\xi \in \Xi$.

The error bound (3.16) (where we can safely set $\kappa = \kappa$, since $\mathbf{Q}_{s,q}(\kappa)$ implies $\mathbf{Q}_{s,1}(\kappa)$) says that (!) holds true for the penalized ℓ_1 recovery with $\lambda = 2s$. The latter observation suggests that the penalized ℓ_1 recovery associated with $(H, \|\cdot\|)$ and $\lambda = 2s$ is better than its regular counterpart, the reason being twofold. First, in order to ensure (!) with the regular recovery, the “built in” parameter ρ of this recovery should be set to $\rho_\epsilon[H, \|\cdot\|]$, and the latter quantity not always is easy to identify. In contrast to this, the construction of penalized ℓ_1 recovery is completely independent of a priori assumptions on the structure of observation errors, while automatically ensuring (!) for the error model we use. Second, and more importantly, for the penalized recovery the bound (3.19) is no more than the “worst, with confidence $1 - \epsilon$, case,” while the typical values of the quantity $\|H^T(u + \xi)\|$ which indeed participates in the error bound (3.15) are essentially smaller than $\rho_\epsilon[H, \|\cdot\|]$. Our numerical experience fully supports the above suggestion: the difference in observed performance of the two routines in question, although not dramatic, is definitely in favour of the penalized recovery. The only potential disadvantage of the latter routine is that the penalty parameter λ should be tuned to the level s of sparsity we aim at, while the regular recovery is free of any guess of this type. Of course, the “tuning” is rather loose – all we need (and experiments show that we indeed need this) is the relation $\lambda \geq 2s$, so that a rough upper bound on s will do; note, however, that the bound (3.15) deteriorates as λ grows.

4 Tractability of condition $\mathbf{Q}_{s,\infty}(\kappa)$, ℓ_∞ -norm of the blocks

We have seen in section 3 that given a sensing matrix A and an r.s. $\mathcal{S} = (B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ such that the associated conditions $\mathbf{Q}_{s,q}(\kappa)$ are satisfiable, we can validate ℓ_1 -recovery of nearly s -block-sparse signals, specifically, can point out ℓ_1 -type recoveries with controlled (and small, provided so are the observation error and the deviation of the signal from an s -block-sparse one). The bad news here is that, in general, condition $\mathbf{Q}_{s,q}(\kappa)$, as well as other conditions for the validity of ℓ_1 recovery, like Block RE or RIP, cannot be verified efficiently. The latter means that given a sensing matrix A and \mathcal{S} , it is difficult to verify that a given candidate pair $(H, \|\cdot\|)$ satisfies the associated with A, \mathcal{S} condition $\mathbf{Q}_{s,q}(\kappa)$. Fortunately, one can construct “tractable approximations” of condition $\mathbf{Q}_{s,q}(\kappa)$, i.e. verifiable sufficient conditions for the

validity of $\mathbf{Q}_{s,q}(\kappa)$. The first good news is that when all $\|\cdot\|_{(k)}$ are the uniform norms $\|\cdot\|_\infty$ and, in addition, $q = \infty$ (which, by Observation 2.1, corresponds to the strongest among the conditions $\mathbf{Q}_{s,q}(\kappa)$ and ensures the validity of (3.12), (3.15) in the largest possible range $1 \leq p \leq \infty$ of values of p), the condition $\mathbf{Q}_{s,q}(\kappa)$ becomes “fully computationally tractable.” We intend to demonstrate also that this condition $\mathbf{Q}_{s,\infty}(\kappa)$ is in fact necessary for the risk bounds of the form (3.12), (3.17) to be valid when $p = \infty$.

4.1 Condition $\mathbf{Q}_{s,\infty}(\kappa)$: tractability and the optimal choice of the contrast H

Notation. In the sequel, given $r, \theta \in [1, \infty]$ and a matrix M , we denote by $\|M\|_{r,\theta}$ the norm of the linear operator $u \mapsto Mu$ induced by the norms $\|\cdot\|_r$ and $\|\cdot\|_\theta$ on the origin and the destination spaces:

$$\|M\|_{r,\theta} = \max_{u: \|u\|_r \leq 1} \|Mu\|_\theta.$$

We denote by $\|M\|_{(\ell,k)}$ the norm of the linear mapping $u \mapsto Mu : \mathbb{R}^{n_\ell} \rightarrow \mathbb{R}^{n_k}$ induced by the norms $\|\cdot\|_{(\ell)}$, $\|\cdot\|_{(k)}$ on the argument and on the image spaces. Further, $\text{Row}_k[M]$ stands for the transpose of the k -th row of M and $\text{Col}_k[M]$ stands for k -th column of M . Finally, $\|u\|_{s,q}$ is the ℓ_q -norm of the vector obtained from a vector $u \in \mathbb{R}^k$ by zeroing all but the s largest entries in u .

Main result. Consider r.s. $\mathcal{S}_\infty = (B, n_1, \dots, n_K, \|\cdot\|_\infty, \dots, \|\cdot\|_\infty)$. We claim that in this case the condition $\mathbf{Q}_{s,\infty}(\kappa)$ becomes fully tractable. Specifically, we have the following

Proposition 4.1 *Let a matrix $A \in \mathbb{R}^{m \times n}$, the r.s. \mathcal{S}_∞ , a positive integer s and reals $\kappa > 0$, $\epsilon \in (0, 1)$ be given.*

(i) *Assume that a triple $(H, \|\cdot\|, \rho)$, where $H \in \mathbb{R}^{m \times M}$, $\|\cdot\|$ is a norm on \mathbb{R}^M , and $\rho \geq 0$, is such that*

$$(!) \ (H, \|\cdot\|) \text{ satisfies } \mathbf{Q}_{s,\infty}(\kappa), \text{ and the set } \Xi = \{\xi : \|H^T[u + \xi]\| \leq \rho \ \forall u \in \mathcal{U}\} \text{ satisfies } P(\Xi) \geq 1 - \epsilon.$$

Given $H, \|\cdot\|, \rho$, one can find efficiently $N = n_1 + \dots + n_K$ vectors h^1, \dots, h^N in \mathbb{R}^m and $N \times N$ block matrix $V = [V^{k\ell}]_{k,\ell=1}^K$ (the blocks $V^{k\ell}$ of V are $n_k \times n_\ell$ matrices) such that

$$\begin{aligned} (a) \quad & B = VB + [h^1, \dots, h^N]^T A, \\ (b) \quad & \|V^{k\ell}\|_{\infty,\infty} \leq s^{-1}\kappa \quad \forall k, \ell \leq K, \\ (c) \quad & P\left(\Xi^+ := \{\xi : \max_{u \in \mathcal{U}} u^T h^i + |\xi^T h^i| \leq \rho, 1 \leq i \leq N\}\right) \geq 1 - \epsilon \end{aligned} \tag{4.20}$$

(note that the matrix norm $\|A\|_{\infty,\infty} = \max_j \|\text{Row}_j[A]\|_1$ is simply the maximal ℓ_1 -norm of the rows of A).

(ii) *Whenever vectors $h^1, \dots, h^N \in \mathbb{R}^m$ and a matrix $V = [V^{k\ell}]_{k,\ell=1}^K$ with $n_k \times n_\ell$ blocks $V^{k\ell}$ satisfy (4.20), the $m \times N$ matrix $\hat{H} = [h^1, \dots, h^N]$, the norm $\|\cdot\|_\infty$ on \mathbb{R}^N and ρ form a triple satisfying (!).*

Discussion. Let a sensing matrix $A \in \mathbb{R}^{m \times n}$ and a r.s. \mathcal{S}_∞ be given, along with a positive integer s , an uncertainty set \mathcal{U} , a distribution P of ξ and $\epsilon \in (0, 1)$. Theorems 3.1, 3.3 say that if a triple $(H, \|\cdot\|, \rho)$ is such that $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,\infty}(\kappa)$ with $\kappa < 1/2$ and H, ρ are such that for the set

$$\Xi = \{\xi : \|H^T[u + \xi]\| \leq \rho \ \forall u \in \mathcal{U}\}$$

it holds $P(\Xi) \geq 1 - \epsilon$, then for the regular ℓ_1 recovery associated with $(H, \|\cdot\|, \rho)$ and for the penalized ℓ_1 recovery associated with $(H, \|\cdot\|)$ and $\lambda = 2s$, the following holds:

$$\begin{aligned} \forall (x \in \mathbb{R}^n, u \in \mathcal{U}, \xi \in \Xi) : \\ L_p(B[\hat{x}(Ax + u + \xi) - x]) \leq \frac{4(2s)^{\frac{1}{p}}}{1-2\kappa} \left[\rho + \frac{1}{2s} L_1(Bx - [Bx]^s) \right], \ 1 \leq p \leq \infty. \end{aligned} \tag{4.21}$$

Proposition 4.1 states that when applying this result, we lose nothing by restricting ourselves with triples $H = [h^1, \dots, h^N] \in \mathbb{R}^{m \times N}$, $N = n_1 + \dots + n_K$, $\|\cdot\| = L_\infty(\cdot)$, $\rho \geq 0$ which can be augmented by an appropriately

chosen matrix $N \times N$ matrix V to satisfy relations (4.20). In the rest of this discussion, it is assumed that we are speaking about triples $(H, \|\cdot\|, \rho)$ satisfying the just defined restrictions.

The bound (4.21) is completely determined by two parameters — κ (which should be $< 1/2$) and ρ ; the smaller are these parameters, the better are the bounds. In what follows we address the issue of efficient synthesis of matrices H with “as good as possible” values of κ and ρ .

Observe first that $H = [h^1, \dots, h^N]$ and κ should admit an extension by a matrix V to a solution of the system of convex constraints (4.20.a), (4.20.b). In the case of $\xi \equiv 0$ the best choice of ρ , given H , is

$$\rho = \max_i \mu_{\mathcal{U}}(h^i), \text{ where } \mu_{\mathcal{U}}(h) = \max_{u \in \mathcal{U}} u^T h.$$

Consequently, in this case the “achievable pairs” ρ, κ form a computationally tractable convex set

$$G_s = \left\{ (\kappa, \rho) : \exists H = [h^1, \dots, h^N] \in \mathbb{R}^{m \times N}, V = [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K : \right. \\ \left. B = VB + H^T A, \|V^{k\ell}\|_{\infty, \infty} \leq \frac{\kappa}{s}, \mu_{\mathcal{U}}(h^i) \leq \rho, 1 \leq i \leq N \right\}.$$

When ξ does not vanish, the situation is complicated by the necessity to maintain the validity of the restriction

$$P(\Xi^+) := P\{\xi : \mu_{\mathcal{U}}(h^i) + |\xi^T h^i| \leq \rho, 1 \leq i \leq N\} \geq 1 - \epsilon, \quad (4.22)$$

which is a chance constraint in variables h^1, \dots, h^N, ρ and as such can be “computationally intractable.” Let us consider the “standard” case of Gaussian zero mean noise ξ , that is, assume that $\xi = D\eta$ with $\eta \sim \mathcal{N}(0, I_m)$ and known $D \in \mathbb{R}^{m \times m}$. Then (4.22) implies that

$$\rho \geq \max_i \left[\mu_{\mathcal{U}}(h^i) + \text{Erfinv}\left(\frac{\epsilon}{2}\right) \|D^T h^i\|_2 \right].$$

On the other hand, (4.22) is clearly implied by

$$\rho \geq \max_i \left[\mu_{\mathcal{U}}(h^i) + \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|D^T h^i\|_2 \right], 1 \leq i \leq N.$$

Ignoring the “gap” between $\text{Erfinv}\left(\frac{\epsilon}{2}\right)$ and $\text{Erfinv}\left(\frac{\epsilon}{2N}\right)$, we can safely model the restriction (4.22) by the system of convex constraints

$$\mu_{\mathcal{U}}(h^i) + \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|D^T h^i\|_2 \leq \rho, 1 \leq i \leq N. \quad (4.23)$$

Thus, the set G_s of admissible κ, ρ can be safely approximated by the computationally tractable convex set

$$G_s^* = \left\{ (\kappa, \rho) : \exists \begin{bmatrix} H = [h^1, \dots, h^N] \in \mathbb{R}^{m \times N} \\ V = [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K \end{bmatrix} : \begin{cases} B = BV + H^T A, \|V^{k\ell}\|_{\infty, \infty} \leq \frac{\kappa}{s}, 1 \leq k, \ell \leq K \\ \max_{u \in \mathcal{U}} u^T h^i + \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|D^T h^i\|_2 \leq \rho, 1 \leq i \leq N \end{cases} \right\}. \quad (4.24)$$

4.2 Condition $\mathbf{Q}_{s,\infty}(\kappa)$: necessity

In this section, as above, we assume that all norms $\|\cdot\|_{(k)}$ in the r.s. \mathcal{S}_∞ are ℓ_∞ -norms; we assume, in addition, that ξ is a zero mean Gaussian noise: $\xi = D\eta$ with $\eta \sim \mathcal{N}(0, I_m)$ and known $D \in \mathbb{R}^{m \times m}$. From the above discussion we know that if, for some $\kappa < 1/2$ and $\rho > 0$, there exist $H = [h^1, \dots, h^N] \in \mathbb{R}^{m \times N}$ and $V = [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K$ satisfying (4.20), then regular and penalized ℓ_1 recoveries with appropriate choice of parameters ensure that

$$\forall (x \in \mathbb{R}^n, u \in \mathcal{U}) : \text{Prob}\{\xi : \|B[x - \hat{x}(Ax + u + \xi)]\|_\infty \leq \frac{4}{1-2\kappa} \left[\rho + \frac{1}{2s} L_1(Bx - [Bx]^s) \right]\} \geq 1 - \epsilon. \quad (4.25)$$

We are about to demonstrate that this implication can be “nearly inverted:”

Proposition 4.2 *Let a sensing matrix A , an r.s. \mathcal{S}_∞ with $\|\cdot\|_{(k)} = \|\cdot\|_\infty$, $1 \leq k \leq K$, an uncertainty set \mathcal{U} , and reals $\kappa > 0$, $\epsilon \in (0, 1/2)$ be given. Suppose that the observation error “is present,” specifically, that for every $r > 0$, the set $\{u + De : u \in \mathcal{U}, \|e\|_2 \leq r\}$ contains a neighborhood of the origin.*

Given a positive integer S , assume that there exists a recovering routine \hat{x} satisfying an error bound of the form (4.25), specifically, the bound

$$\forall (x \in \mathbb{R}^n, u \in \mathcal{U}) : \text{Prob}\{\|B[x - \hat{x}(Ax + u + \xi)]\|_\infty \leq \alpha + S^{-1}L_1(Bx - [Bx]^S)\} \geq 1 - \epsilon. \quad (4.26)$$

for some $\alpha > 0$. Then there exist $H = [h^1, \dots, h^N] \in \mathbb{R}^{m \times N}$ and $V = [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K$ satisfying

$$\begin{aligned} (a) \quad & B = VB + H^T A, \\ (b) \quad & \|V^{k\ell}\|_{\infty, \infty} \leq 2S^{-1} \quad \forall k, \ell \leq K, \\ (c) \quad & \text{with } \rho := \max_{1 \leq i \leq N} \left[\max_{u \in \mathcal{U}} u^T h^i + \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|D^T h^i\|_2 \right], \end{aligned} \quad (4.27)$$

one has $\rho \leq 2\alpha$ when $D = 0$, $\rho \leq 2\alpha \frac{\text{Erfinv}(\frac{\epsilon}{2N})}{\text{Erfinv}(\epsilon)}$ when $D \neq 0$, and for $\xi = D\eta$, $\eta \sim \mathcal{N}(0, I_m)$ one has

$$P\left(\Xi^+ := \{\xi : \max_{u \in \mathcal{U}} u^T h^i + |\xi^T h^i| \leq \rho, 1 \leq i \leq N\}\right) \geq 1 - \epsilon.$$

In other words, (see Proposition 4.1), $(H, L_\infty(\cdot))$ satisfies $\mathbf{Q}_{s,\infty}(\kappa)$ for s “nearly as large as S ,” namely, $s \leq \frac{\kappa}{2}S$, and $H = [h^1, \dots, h^N]$, ρ satisfy conditions (4.23) with ρ being “nearly α ”, namely, $\rho \leq 2\alpha$ in the case of $D = 0$ and $\rho \leq 2 \frac{\text{Erfinv}(\frac{\epsilon}{2N})}{\text{Erfinv}(\epsilon)}$ when $D \neq 0$. In particular, under the premise of Proposition 4.2, the contrast optimization procedure of section 4.1 supplies the matrix H such that the corresponding regular or penalized recovery $\hat{x}(\cdot)$ for all $s \leq \frac{S}{8}$ satisfies:

$$\text{Prob}\left\{\|B[x - \hat{x}(y)]\|_\infty \leq 4 \left[4 \frac{\text{Erfinv}(\frac{\epsilon}{2N})}{\text{Erfinv}(\epsilon)} \alpha + s^{-1}L_1(Bx - [Bx]^s)\right]\right\} \geq 1 - \epsilon.$$

5 Tractable approximations of $\mathbf{Q}_{s,q}(\kappa)$

Aside from the important case $q = \infty$, $\|\cdot\|_{(k)} = \|\cdot\|_\infty$ considered in sections 4.1 and 4.2, condition $\mathbf{Q}_{s,q}(\kappa)$ “as it is” seems to be computationally intractable: unless $s = O(1)$, it is unknown how to check efficiently that a given pair $(H, \|\cdot\|)$ satisfies this condition, not speaking about synthesis of a pair satisfying this condition and resulting in the best possible error bound 3.12), (3.15) for regular and penalized ℓ_1 -recoveries. We are about to present *verifiable sufficient conditions* for the validity of $\mathbf{Q}_{s,q}(\kappa)$ which may become and interesting substitution for condition $\mathbf{Q}_{s,q}(\kappa)$ for that purposes.

5.1 Sufficient condition for $\mathbf{Q}_{s,q}(\kappa)$

Proposition 5.1 *Suppose that a sensing matrix A , an r.s. $(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$, and $\kappa \geq 0$ are given.*

Let $N = n_1 + \dots + n_K$, and let $N \times N$ matrix $V = [V^{k\ell}]_{k,\ell=1}^K$ ($V^{k\ell}$ are $n_k \times n_\ell$) and $m \times N$ matrix H satisfy the relation

$$B = VB + H^T A. \quad (5.28)$$

Let us denote

$$\nu_{s,q}^*(V) = \max_{1 \leq \ell \leq K} \max_{w^\ell \in \mathbb{R}^{n_\ell} : \|w^\ell\|_{(\ell)} \leq 1} L_{s,q}\left([V^{1\ell}w^\ell; \dots; V^{K\ell}w^\ell]\right).$$

Then for all $s \leq K$ and all $q \in [1, \infty]$, we have:

$$L_{s,q}(Bx) \leq s^{\frac{1}{q}} L_\infty(H^T Ax) + \nu_{s,q}^*(V) L_1(Bx) \quad \forall x. \quad (5.29)$$

The result of Proposition 5.1 is a step to verifiable sufficient condition for the validity of $Q_{s,q}$. To get such condition we need an efficiently computable upper bound of the quantity $\nu_{s,q}^*$. In particular, if for a given positive integer $s \leq K$ and a real $q \in [1, \infty]$ there exist an upper bounding function $\nu_{s,q}(V)$ such that

$$\nu_{s,q}(\cdot) \text{ is convex and } \nu_{s,q}(V) \geq \nu_{s,q}^*(V) \forall V \quad (5.30)$$

and a matrix V such that

$$\nu_{s,q}(V) \leq s^{\frac{1}{q}-1} \kappa, \quad (5.31)$$

then the pair $(H, L_\infty(\cdot))$ satisfies $\mathbf{Q}_{s,q}(\kappa)$. An important example of the upper bound for $\nu_{s,q}^*(V)$ which satisfies (5.31) is provided in the following statement.

Proposition 5.2 *Let Ω be a $K \times K$ matrix with entries $[\Omega]_{k,\ell} = \|V^{k\ell}\|_{(\ell,k)}$, $1 \leq k, \ell \leq K$. Then*

$$\widehat{\nu}_{s,q}(V) := \max_{1 \leq k \leq K} \|\text{Col}_k[\Omega]\|_{s,q} \geq \nu_{s,q}^*(V) \forall V \quad (5.32)$$

(note that the inequality in (5.32) becomes equality when either $q = \infty$, or $s = 1$), so that the condition

$$\widehat{\nu}_{s,q}(V) \leq s^{\frac{1}{q}-1} \kappa \quad (5.33)$$

is sufficient for $(H, L_\infty(\cdot))$ to satisfy $\mathbf{Q}_{s,q}(\kappa)$.

When all $\|\cdot\|_{(k)}$ are the ℓ_∞ -norms and $q = \infty$, the results of Propositions 5.1 and 5.2 recover Proposition 4.1. In the general case, they suggest a way to synthesize matrices $H \in \mathbb{R}^{m \times N}$ which, taken along with the norm $\|\cdot\| = L_\infty(\cdot)$, provably satisfy the condition $\mathbf{Q}_{s,q}(\kappa)$, along with a certificate V for this fact. Namely, H and V should satisfy the system of linear equations (5.28) and, in addition, (5.31) should hold for V with $\nu_{s,q}(\cdot)$ satisfying (5.30). Further, for such a $\nu_{s,q}(\cdot)$, (5.31) is a system of convex constraints on V . Whenever these constraints are efficiently computable, we get a computationally tractable sufficient condition on H to satisfy $\mathbf{Q}_{s,q}(\kappa)$ – a condition which is expressed by an explicit system of efficiently computable convex constraints (5.28), (5.31) on H and additional matrix variable V .

5.2 Tractable sufficient conditions and contrast optimization

The quantity $\widehat{\nu}_{s,q}(\cdot)$ is the simplest choice of $\nu_{s,q}(\cdot)$ satisfying (5.30). In this case, efficient computability of the constraints (5.31) is the same as efficient computability of norms $\|\cdot\|_{(k,\ell)}$. Assuming that $\|\cdot\|_{(k)} = \|\cdot\|_{r_k}$ for every k , the computability issue becomes the one of efficient computation of the norms $\|\cdot\|_{r_\ell, r_k}$. The norm $\|\cdot\|_{r,\theta}$ is known to be generically efficiently computable in only three cases:

1. $\theta = \infty$, where $\|M\|_{r,\infty} = \|M^T\|_{1,\frac{r}{r-1}} = \max_i \|\text{Row}_i^T(M)\|_{\frac{r}{r-1}}$;
2. $r = 1$, where $\|M\|_{1,\theta} = \max_j \|\text{Col}_j[M]\|_\theta$;
3. $r = \theta = 2$, where $\|M\|_{2,2} = \sigma_{\max}(M)$ is the spectral norm of M .

Assuming for the sake of simplicity that in our r.s. $\|\cdot\|_{(k)}$ are r -norms with common value of r , let us look at three “tractable cases” as specified by the above discussion – those of $r = \infty$, $r = 1$ and $r = 2$. In these cases, candidate contrast matrices H are $m \times N$, the associated norm $\|\cdot\|$ is $L_\infty(\cdot)$, and our sufficient condition for H to be good (i.e., for $(H, L_\infty(\cdot))$ to satisfy $\mathbf{Q}_{s,q}(\kappa)$ with given $\kappa < 1/2$ and q) becomes a system $\mathbf{S} = \mathbf{S}_{\kappa,q}$ of explicit efficiently computable convex constraints on H and additional matrix variable $V \in \mathbb{R}^{N \times N}$, implying that the set \mathcal{H} of good H is convex and computationally tractable, so that we can minimize efficiently over \mathcal{H} any convex and efficiently computable function. In our context, a natural way to use \mathbf{S} is to optimize over $H \in \mathcal{H}$ the error bound (3.19), or, which is the same, to minimize over \mathcal{H} the function $\rho(H) = \rho_\epsilon[H, L_\infty(\cdot)]$, see (3.18), where $\epsilon < 1$ is a given tolerance. Taken literally, this problem

still can be difficult, since the function $\rho(H)$ is not necessarily convex, and can be difficult to compute even when convex. To overcome this difficulty, we again can use a *verifiable sufficient condition* for the relation $\rho(H) \leq \rho$, that is, a system $\mathbf{T} = \mathbf{T}_\epsilon$ of explicit efficiently computable convex constraints on variables H and ρ (and, perhaps some slack variables ζ) such that $\rho(H) \leq \rho$ for the (H, ρ) -component of every feasible solution of \mathbf{T} . With this approach, design of the best, as allowed by \mathbf{S} and \mathbf{T} , contrast matrix H reduces to solving a convex optimization problem with efficiently computable constraints in variables H, V, ρ , specifically, the problem

$$\min_{\rho, H, V, \zeta} \{ \rho : H, V \text{ satisfy } \mathbf{S}; H, \rho, \zeta \text{ satisfy } \mathbf{T} \}. \quad (5.34)$$

In the rest of this section, we present explicitly the systems \mathbf{S} and \mathbf{R} for the three tractable cases we are interested in, assuming the following model of observation errors:

$$\mathcal{U} = \{u = Ev : \|v\|_2 \leq 1\}; \quad \xi = D\eta, \quad \eta \sim \mathcal{N}(0, I_m),$$

where $E, D \in \mathbb{R}^{m \times m}$.

We use the following notation: the $m \times n$ matrix H is partitioned into $m \times n_k$ blocks $H[k]$, $1 \leq k \leq K$, according to the block structure of the representation vectors; t -th column in $H[k]$ is denoted $h^{kt} \in \mathbb{R}^m$, $1 \leq t \leq n_k$.

For derivations of the results to follow, see section A.7.

The case of $r = \infty$. The case of $q = \infty$ was considered in full details in section 4.1. When $q \leq \infty$, one has:

$$\begin{aligned} \mathbf{S}_{\kappa, q} : & \begin{cases} B = VB + H^T A, \quad \Omega_{k\ell} := \|V^{k\ell}\|_{\infty, \infty} = \max_{1 \leq t \leq n_k} \|\text{Row}_t[V^{k\ell}]\|_1, \quad 1 \leq k, \ell \leq K, \\ \|\text{Col}_\ell[\Omega]\|_{s, q} \leq s^{\frac{1}{q}-1} \kappa, \quad 1 \leq \ell \leq K \end{cases} \\ \mathbf{R}_\epsilon : & \text{Erfinv}\left(\frac{\epsilon}{2N}\right) \|D^T h^{kt}\|_2 + \|E^T h^{kt}\|_2 \leq \rho, \quad 1 \leq t \leq n_k, 1 \leq k \leq K. \end{aligned} \quad (5.35)$$

The case of $r = 2$. Here

$$\begin{aligned} \mathbf{S}_{\kappa, q} : & \begin{cases} B = VB + H^T A, \quad \Omega_{k\ell} := \|V^{k\ell}\|_{2,2} = \sigma_{\max}(V^{k\ell}), \quad 1 \leq k, \ell \leq K \\ \|\text{Col}_\ell[\Omega]\|_{s, q} \leq s^{\frac{1}{q}-1} \kappa, \quad 1 \leq \ell \leq K \end{cases} \\ \mathbf{R}_\epsilon : & \left. \begin{aligned} & \exists \{W_k \in \mathbb{S}^m, \alpha_k, \beta_k, \gamma_k \in \mathbb{R}\}_{k=1}^K : \\ & \sigma_{\max}(E^T H[k]) + \alpha_k \leq \rho \\ & \begin{bmatrix} W_k & D^T H[k] \\ H^T[k] D & \alpha_k I_{n_k} \end{bmatrix} \succeq 0, \quad \|\lambda(W_k)\|_\infty \leq \beta + k, \quad \|\lambda(W_k)\|_2 \leq \gamma_k \\ & \text{Tr}(W_k) + 2 \left[\delta \beta_k + \sqrt{\delta^2 \beta_k^2 + 2\delta \gamma_k^2} \right] \leq \alpha_k \end{aligned} \right\}, \quad 1 \leq k \leq K \\ & \delta := \ln(K/\epsilon), \end{aligned} \quad (5.36)$$

where \mathbb{S}^m is the space of $m \times m$ symmetric matrices, $\lambda(W)$ is the vector of eigenvalues of $W \in \mathbb{S}^m$.

The case of $r = 1$. Here

$$\begin{aligned} \mathbf{S}_{\kappa, q} : & \begin{cases} B = VB + H^T A, \quad \Omega_{k\ell} := \|V^{k\ell}\|_{1,1} = \max_{1 \leq t \leq n_\ell} \|\text{Col}_t[V^{k\ell}]\|_1, \quad 1 \leq k, \ell \leq K, \\ \|\text{Col}_\ell[\Omega]\|_{s, q} \leq s^{\frac{1}{q}-1} \kappa, \quad 1 \leq \ell \leq K \end{cases} \\ \mathbf{R}_\epsilon : & \left. \begin{aligned} & \exists \{\lambda^k \in \mathbb{R}_+^m, \mu^k \geq 0\}_{k=1}^K : \\ & \text{Erfinv}\left(\frac{\epsilon}{2Kn_k}\right) \sum_{t=1}^{n_k} \|D^T h^{kt}\|_2 + \frac{1}{2} \sum_i \lambda_i^k + \frac{1}{2} \mu^k \leq \rho \\ & \begin{bmatrix} \text{Diag}\{\lambda^k\} & H^T[K] E \\ E^T H[k] & \mu^k I_{n_k} \end{bmatrix} \succeq 0 \end{aligned} \right\} \quad \forall (k \leq K, t \leq n_k). \end{aligned} \quad (5.37)$$

5.3 Tractable sufficient conditions: limits of performance

Consider the situation where all the norms $\|\cdot\|_{(k)}$ are $\|\cdot\|_r$, with $r \in \{1, 2, \infty\}$. A natural question about verifiable sufficient conditions for a pair $(H, L_\infty(\cdot))$ to satisfy $\mathbf{Q}_{s,q}(\kappa)$ is, what are the “limits of performance” of these sufficient conditions. Specifically, how large could be the range of s for which the condition can be satisfied by at least one contrast matrix. Here is a partial answer to this question:

Proposition 5.3 *Let A be an $m \times n$ sensing matrix which is “essentially non-square,” specifically, such that $2m \leq n$, let $B = I$ and let $n_k = d$, $\|\cdot\|_{(k)} = \|\cdot\|_r$, $1 \leq k \leq K$, with $r \in \{1, 2, \infty\}$. Whenever an $m \times N$ matrix H and $N \times N$ matrix V satisfy the conditions*

$$I = V + H^T A \text{ and } \max_{1 \leq \ell \leq K} \|\|V^{1\ell}\|_{r,r}; \|V^{2\ell}\|_{r,r}; \dots; \|V^{K\ell}\|_{r,r}\|_{s,q} \leq \frac{1}{2} s^{\frac{1}{q}-1} \quad (5.38)$$

(cf. (5.28), (5.32), and (5.31)), one has

$$s \leq \frac{3\sqrt{m}}{2\sqrt{d}}. \quad (5.39)$$

Discussion. Let the r.s. in question be the same as in Proposition 5.3, and let $m \times n$ sensing matrix A have $2m \leq n$. Proposition 5.3 says that in this case, verifiable sufficient condition, stated by Proposition 5.1, for satisfiability of $\mathbf{Q}_{s,q}(\kappa)$ with $\kappa < 1/2$ has rather restricted scope — it cannot certify the satisfiability of $\mathbf{Q}_{s,q}(\kappa)$, $\kappa \leq 1/2$, when $s \geq \frac{3\sqrt{m}}{2\sqrt{d}}$. Yet, the condition $\mathbf{Q}_{s,q}(\kappa)$ may be satisfiable in a much larger range of values of s . For instance, when the r.s. in question is the standard one, and A is a random Gaussian $m \times n$ matrix, the matrix A satisfies, with overwhelming probability as m, n grow, the $\text{RIP}(\frac{1}{5}, s)$ condition for s as large as $O(1)m/\sqrt{\ln(n/m)}$ (cf. [8]). By Proposition 2.1, this implies that $(\frac{5}{4}A, \|\cdot\|_\infty)$ satisfies the condition $\mathbf{Q}_{s,2}(\frac{1}{4})$ in the essentially the same large range of s . There is, however, an important case where the “limits of performance” of our verifiable sufficient condition for the satisfiability of $\mathbf{Q}_{s,q}(\kappa)$ implies severe restrictions on the range of values of s in which the “true” condition $\mathbf{Q}_{s,q}(\kappa)$ is satisfiable — this is the case when $q = \infty$ and $r = \infty$. Combining Propositions 4.1 and 5.3, we conclude that in the case of r.s. from Proposition 5.3 with $r = \infty$ and “sufficiently non-square” ($2m \leq n$) $m \times n$ sensing matrix A , the associated condition $\mathbf{Q}_{s,\infty}(\frac{1}{2})$ is cannot be satisfied when $s > \frac{3\sqrt{n}}{2\sqrt{d}}$.

5.4 Tractable sufficient conditions and Mutual Block-Incoherence

We have mentioned in Introduction that, to the best of our knowledge, the only previously proposed verifiable sufficient condition for the validity of block- ℓ_1 recovery is the “mutual block incoherence condition” [15]. Our immediate goal is to show that this condition is covered by Proposition 5.1.

Consider an r.s. with $B = I$ and with ℓ_2 -norms in the role of $\|\cdot\|_{(k)}$, $1 \leq k \leq K$, and let the sensing matrix A in question be partitioned as $A = [A[1], \dots, A[K]]$, where $A[k]$ has n_k columns. Let us define the mutual block-incoherence μ of A w.r.t. the r.s. in question as follows:

$$\mu = \max_{\substack{1 \leq k, \ell \leq K, \\ k \neq \ell}} \sigma_{\max}(C_k^{-1} A^T[k] A[\ell]), \quad [C_k := A^T[k] A[k]] \quad (5.40)$$

provided that all matrices C_k , $1 \leq k \leq K$, are nonsingular, otherwise $\mu = \infty$. Note that in the case of the standard r.s., the just defined quantity is nothing but the standard mutual incoherence known from the Compressed Sensing literature (see, e.g., [13]).

In [15], the authors consider the same r.s. and assume that $n_k = d$, $1 \leq k \leq K$, and that the columns of A are of unit $\|\cdot\|_2$ -norm. They introduce the quantities

$$\nu = \max_{1 \leq k \leq K} \max_{1 \leq j \neq j' \leq K} |\text{Col}_j^T[A[k]] \text{Col}_{j'}[A[k]]|, \quad \mu_B = \frac{1}{d} \max_{\substack{1 \leq k, \ell \leq K, \\ k \neq \ell}} \sigma_{\max}(A^T[k] A[\ell]) \quad (5.41)$$

and prove that an appropriate version of block- ℓ_1 recovery allows to recover exactly every s -block-sparse signal x from the noiseless observations $y = Ax$, provided that

$$1 - (d-1)\nu > 0 \text{ and } s < \chi := \frac{1 - (d-1)\nu + d\mu_B}{2d\mu_B}. \quad (5.42)$$

The following observation is almost immediate:

Proposition 5.4 *Given $m \times n$ sensing matrix A and an r.s. \mathcal{S} with $B = I$, $\|\cdot\|_{(k)} = \|\cdot\|_2$, $1 \leq k \leq K$, let $A = [A[1], \dots, A[K]]$ be the corresponding partition of A .*

(i) *Let μ be the mutual block-incoherence of A w.r.t. \mathcal{S} . Assuming $\mu < \infty$, we set*

$$H = \frac{1}{1+\mu} [A[1]C_1^{-1}, A[2]C_2^{-1}, \dots, A[K]C_K^{-1}], \quad C_k = A^T[k]A[k]. \quad (5.43)$$

Then the contrast matrix H along with the matrix $V = I - H^T A$ satisfies condition (5.28) (where $B = I$) and condition (5.33) with $q = \infty$ and

$$\kappa = \frac{\mu s}{1+\mu}.$$

As a result, applying Proposition 5.1, we conclude that whenever

$$s < \frac{1+\mu}{2\mu}, \quad (5.44)$$

the pair $(H, L_\infty(\cdot))$ satisfies $\mathbf{Q}_{s,\infty}(\kappa)$ with $\kappa = \frac{\mu s}{1+\mu} < 1/2$.

(ii) *Suppose that $n_k = d$, $k = 1, \dots, K$, and let the quantities ν and μ_B defined in (5.41) satisfy the relations (5.42). Then the mutual block-incoherence of A w.r.t. the r.s. in question does not exceed $\bar{\mu} = \frac{d\mu_B}{1-(d-1)\nu}$. Further, we have $\frac{1+\bar{\mu}}{2\bar{\mu}} = \chi$, and (5.44) holds, and thus ensures that the contrast H , as defined in (5.43), and $L_\infty(\cdot)$ satisfy $\mathbf{Q}_{s,\infty}(\kappa)$ with some $\kappa < \frac{1}{2}$.*

Let $A = [A_{ij}] \in \mathbb{R}^{m \times n}$ be random matrix with i.i.d. entries $A_{ij} \sim cN(0, m^{-1})$. We have the following simple result.

Proposition 5.5 *Assume that $B = I$, $n_k = d$ and $\|\cdot\|_{(k)} = \|\cdot\|_2$ for all k . There are absolute constants $C_1, C_2 < \infty$ (the corresponding bounds are provided in the appendix A.10) such that if $m \geq C_1(d + \ln(n))$, then the mutual block-incoherence μ of A satisfies with probability at least $1 - \frac{1}{n}$*

$$\mu \leq C_2 \sqrt{\frac{d + \ln(n)}{m}}. \quad (5.45)$$

The bound (5.45), along with Proposition 5.4.(i), implies that when A is a Gaussian matrix, all block-norms are the ℓ_2 -norms and all $n_k = d$ with d “large enough” (such that $d^{-1} \ln n = O(1)$), the verifiable sufficient condition for $\mathbf{Q}_{s,\infty}(\frac{1}{3})$ holds with overwhelming probability for $s = O(\sqrt{\frac{m}{d}})$. In other words, in this case the (verifiable!) condition $\mathbf{Q}_{s,\infty}(\kappa)$ attains (up to an absolute factor) the limit of performance stated in Proposition 5.3.

6 Matching pursuit algorithm for block recovery

The Matching Pursuit algorithm for block-sparse recovery is motivated by the desire to provide a reduced complexity alternative to the algorithms using ℓ_1 -minimization. Several implementations of Matching Pursuit for block-sparse recovery have been proposed in the Compressed Sensing literature [3, 4, 15, 16]. In this section we aim to show that a pair H, V satisfying (5.28) and (5.31) where $\kappa < 1/2$ (and thus, by Proposition 5.1, such that $(H, L_\infty(\cdot))$ satisfies $\mathbf{Q}_{s,\infty}(\kappa)$) can be used to design a specific version of the Matching Pursuit

algorithm which we refer to as *Non-Euclidean Block Matching Pursuit (NEBMP) algorithm* for block-sparse recovery.

We fix an r.s. $\mathcal{S} = (B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ and assume that the block norms $\|\cdot\|_{(k)}$, $k = 1, \dots, K$, are either $\|\cdot\|_\infty$ - or $\|\cdot\|_2$ -norms. Furthermore, we suppose that the matrix B is of full row rank, so that given $z \in \mathbb{R}^N$ one can compute x such that $z = Bx$ (e.g., $x = B^+ z$ where $B^+ = B^T(BB^T)^{-1}$ is the pseudo-inverse of B). Let the noise ξ in the observation $y = Ax + u + \xi$ be Gaussian, $\xi \sim \mathcal{N}(0, D)$, $D \in \mathbb{R}^{m \times m}$ is known. Finally, we assume that we are in the situation of section 5.2, that is, we have at our disposal an $m \times N$, $N = n_1 + \dots + n_K$, matrix H , an $N \times N$ block matrix $V = [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K$, a $\bar{\gamma} > 0$ and $\rho \geq 0$ such that

$$\begin{aligned} (a) \quad & B = VB + H^T A, \\ (b) \quad & \|V^{k\ell}\|_{(\ell,k)} \leq \bar{\gamma} \quad \forall k, \ell \leq K \\ (c) \quad & \text{Prob}\{\Xi^+ := \{\xi : L_\infty(H^T[u + \xi]) \leq \rho \forall u \in \mathcal{U}\}\} \geq 1 - \epsilon. \end{aligned} \tag{6.46}$$

Given observation y , a positive integer s and a real $v \geq 0$ (v is our guess for an upper bound on $L_1(Bx - [Bx]^s)$), consider Algorithm 1 below. Its convergence analysis is based upon the following:

Algorithm 1 Non-Euclidean Block Matching Pursuit

1. Initialization: Set $v^{(0)} = 0$, $\alpha_0 = \frac{L_{s,1}(H^T y) + s\rho + v}{1 - s\bar{\gamma}}$.

2. Step k , $k = 1, 2, \dots$: Given $v^{(k-1)} \in \mathbb{R}^n$ and $\alpha_{k-1} \geq 0$, compute

(a) $g = H^T(y - Av^{(k-1)})$ and vector $\Delta = [\Delta[1], \dots, \Delta[K]] \in \mathbb{R}^N$ by setting for $j = 1, \dots, K$:

$$\begin{aligned} \Delta[j] &= \frac{g[j]}{\|g[j]\|_2} [\|g[j]\|_2 - \bar{\gamma}\alpha_{k-1} - \nu(H)]_+, \text{ if } \|\cdot\|_{(j)} = \|\cdot\|_2; \\ \Delta_{ji} &= \text{sign}(g_{ji}) [|g_{ji}| - \bar{\gamma}\alpha_{k-1} - \nu(H)]_+, 1 \leq i \leq n_j, \text{ if } \|\cdot\|_{(j)} = \|\cdot\|_\infty, \end{aligned} \tag{6.47}$$

where w_{ji} is i -th entry in j -th block of a representation vector w .

(b) Choose $v^{(k)}$ such that $B(v^{(k)} - v^{(k-1)}) = \Delta$, set

$$\alpha_k = 2s\bar{\gamma}\alpha_{k-1} + 2s\rho + v. \tag{6.48}$$

and loop to step $k + 1$.

3. Output: the approximate solution found after k iterations is $v^{(k)}$.

Lemma 6.1 *In the situation of (6.46), let $s\bar{\gamma} < 1$. Then whenever $\xi \in \Xi^+$, for every $x \in \mathbb{R}^n$ with $L_1(Bx - [Bx]^s) \leq v$ and every $u \in \mathcal{U}$, the following holds true.*

When applying Algorithm 1 to $y = Ax + u + \xi$, the resulting approximations $Bv^{(k)}$ to Bx and the quantities α_k for all k satisfy the relations

$$\begin{aligned} (a_k) \quad & \text{for all } 1 \leq j \leq K \quad \|(Bv^{(k)} - Bx)[j]\|_{(j)} \leq \|(Bx)[j]\|_{(j)}, \\ (b_k) \quad & L_1(Bx - Bv^{(k)}) \leq \alpha_k \text{ and } L_\infty(Bx - Bv^{(k+1)}) \leq 2\bar{\gamma}\alpha_k + 2\rho. \end{aligned}$$

Note that if $2s\bar{\gamma} < 1$, then also $s\bar{\gamma} < 1$, so that Lemma 6.1 is applicable. Furthermore, in this case, by (6.48), the sequence α_k converges exponentially fast to the limit $\alpha_\infty := \frac{2s\rho + v}{1 - 2s\bar{\gamma}}$:

$$L_1(Bv^{(k)} - Bx) \leq \alpha_k = (2s\bar{\gamma})^k [\alpha_0 - \alpha_\infty] + \alpha_\infty.$$

Along with the second inequality of (b_k) this implies the bounds:

$$L_\infty(Bv^{(k)} - Bx) \leq 2\bar{\gamma}\alpha_{k-1} + 2\rho \leq \frac{\alpha_k}{s},$$

and since $L_p(w) \leq L_1(w)^{\frac{1}{p}} L_\infty(w)^{\frac{p-1}{p}}$ for $1 \leq p \leq \infty$, we have

$$L_p(Bv^{(k)} - Bx) \leq s^{\frac{1-p}{p}} \left[(2s\bar{\gamma})^k [\alpha_0 - \alpha_\infty] + \alpha_\infty \right].$$

The bottom line here is as follows.

Proposition 6.1 *Suppose that a collection $(H, L_\infty(\cdot), \rho, \bar{\gamma}, \epsilon)$ satisfies (6.46), and let the parameter s of Algorithm 1 satisfy $\kappa := 2s\bar{\gamma} < 1$. Then for all $\xi \in \Xi^+$, $u \in \mathcal{U}$, $x \in \mathbb{R}^n$ such that $L_1(Bx - [Bx]^s) \leq v$, Algorithm 1 as applied to $y = Ax + u + \xi$ ensures that for every $t = 1, 2, \dots$ one has*

$$L_p(Bv^{(t)} - Bx) \leq s^{\frac{1}{p}} \left[\frac{2\rho + s^{-1}v}{1 - 2\kappa} + (2\kappa)^t \left[\frac{s^{-1}(L_{s,1}(H^T y) + v) + \rho}{1 - \kappa} - \frac{2\rho + s^{-1}v}{1 - 2\kappa} \right] \right], \quad 1 \leq p \leq \infty$$

(cf. (4.21)).

Note that Proposition 6.1 combined with Proposition 5.4 essentially covers the results of [15] on the properties of the Matching Pursuit algorithm for block-sparse recovery proposed in this reference.

7 Numerical Illustration

In the theoretical part of this paper we were looking at the situation where the sensing matrix A and the r.s. $(B, n_1, \dots, n_K, \|\cdot\|_{(1)}, \dots, \|\cdot\|_{(K)})$ were given, and we were interested to understand

[A] whether ℓ_1 recovery allows to recover the representations Bx of all s -block-sparse signals with a given s in the absence of observation noise, and

[B] how to choose the best (resulting in the smallest possible error bounds) pair $(H, \|\cdot\|)$.³

Note that different components of our setup have in fact different status. While in typical applications A , B and the block structure n_1, \dots, n_K of the representation vectors may be thought as conditioned by the “problem’s physics”, it is not the case for the block norms $\|\cdot\|_{(k)}$. Their choice (which does affect the ℓ_1 recovery routines) appears to be unrelated to the model of the data.

The first goal of our experiments is to understand how to choose the block norms in order to validate ℓ_1 recovery for the largest possible value of the sparsity parameter s . The meaning we put into the word “validate” is providing guarantees of small error of recovery of all s -block-sparse signals when the observation error is small (what implies, of course, the exactness of the recovery in the case of noiseless observation). Here we restrict ourselves to the case when all block norms are ℓ_r norms with common value of r taken from the range $\{1, 2, \infty\}$. By reasons explained in the discussion in section 3, we consider here only the case of the penalized ℓ_1 -recovery with $m \times N$ contrast matrix H (where, as always, $N = n_1 + \dots + n_K$), $\|\cdot\| = L_\infty(\cdot)$ ⁴, and with $\lambda = 2s$ (see (3.14)). Beside this, we assume, mainly for the sake of notational convenience, that $B = I$.

Let us fix $A \in \mathbb{R}^{m \times n}$, $B = I_n$, K, n_1, \dots, n_K ($n_1 + \dots + n_K = n =: N$). By Proposition 5.1, for every matrix $H \in \mathbb{R}^{m \times n}$ setting

$$\begin{aligned} V &\equiv [V^{k\ell} \in \mathbb{R}^{n_k \times n_\ell}]_{k,\ell=1}^K = I - H^T A, \quad \Omega^r(H) = [\|V^{k\ell}\|_{r,r}]_{k,\ell=1}^K, \\ \kappa_1^{r,s}(H) &= \max_{1 \leq \ell \leq K} \|\text{Col}_\ell[\Omega^r(H)]\|_{s,1}, \quad \kappa_\infty^{r,s}(H) = s \max_{1 \leq k, \ell \leq K} [\Omega^r(H)]_{k,\ell}, \end{aligned} \quad (7.49)$$

³Needless to say, the results presented so far do not pretend to provide full answers to these questions. Our verifiable sufficient conditions for the “validity” of block- ℓ_1 recovery supply only some *lower bounds* on the largest $s = s_*$ for which the answer to [A] is positive. Similarly, aside of the case $q = \infty$, $\|\cdot\|_{(k)} = \|\cdot\|_\infty$, $1 \leq k \leq K$, our conditions for the validity of block- ℓ_1 recovery are only sufficient, meaning that optimizing the error bound over $(H, \|\cdot\|)$ allowed by these conditions may only yield *suboptimal* recovery routines.

⁴these are exactly the pairs $(H, \|\cdot\|)$ covered by the sufficient conditions for the validity of ℓ_1 -recovery, see Proposition 5.1.

the pair $(H, L_\infty(\cdot))$ satisfies the conditions $\mathbf{Q}_{s,q}(\kappa_q^{r,s}(H))$, $q = 1$ and $q = \infty$, provided that the block norms are the ℓ_r -ones. In particular, when $\kappa_1^{r,s}(H) < 1/2$, the penalized ℓ_1/ℓ_r recovery (i.e. the recovery (3.10) with all block norms being the ℓ_r -ones) “is valid” on s -block-sparse signals, meaning exactly that this recovery ensures the validity of the error bounds (3.17) with $q = \infty$, $\varkappa = \kappa_1^{r,s}$, $\kappa = \kappa_\infty^{r,s}$ (and, in particular, recovers exactly all s -block-sparse signals when there is no observation noise).

Our strategy is as follows. For each value of $r \in \{1, 2, \infty\}$, we consider the convex optimization problem

$$\min_{H \in \mathbb{R}^{m \times n}} \left\{ \kappa_1^{r,s}(H) := \max_{\ell \leq K} \|\text{Col}_\ell[\Omega^r(H)]\|_{s,1} \right\},$$

find the largest $s = s(r)$ for which the optimal value in this problem is $< 1/2$, and denote by $H^{(r)}$, $r \in \{1, 2, \infty\}$ the corresponding optimal solution. To these “marked” contrast matrices we add two more contrasts, $H^{(\text{MI})}$ and $H^{(\text{MBI})}$, based on the mutual block-incoherence condition and given by the calculation (5.40) for the cases of the “standard” (1-element blocks in $x = Bx$) and the actual block structures, respectively.

Now, given the set $\mathcal{H} = \{H^{(\text{MI})}, H^{(\text{MBI})}, H^{(1)}, H^{(2)}, H^{(\infty)}\}$ of $m \times n$ candidate contrast matrices, we can choose the “most powerful” penalized ℓ_1/ℓ_r recovery suggested by \mathcal{H} as follows: for every $H \in \mathcal{H}$ and for every $p \in \{1, 2, \infty\}$, we find the largest $s = s(H, p)$ for which $\kappa_1^{r,p}(H) < 1/2$, and then define the quantity $s_* = s_*(\mathcal{H}) = \max\{s(H, p) : H \in \mathcal{H}, p \in \{1, 2, \infty\}\}$ along with $H_* \in \mathcal{H}$ and $p_* \in \{1, 2, \infty\}$ such that $s_* = s(H_*, p_*)$. The penalized ℓ_1/ℓ_{p_*} recovery utilizing the contrast matrix H_* and the norm $L_\infty(\cdot)$ associated with block norms $\|\cdot\|_{p_*}$ of the blocks is definitely valid for $s = s_*(\mathcal{H})$, and this is the largest sparsity range, as certified by our sufficient conditions for the validity of ℓ_1/ℓ_r recovery, which we can get with contrast matrices from \mathcal{H} . Note that $s_* \geq \max[s(1), s(2), s(\infty)]$, that is, the resulting range of values of s is also the largest we can certify using our sufficient conditions, with *no* restriction on the contrast matrices.

Implementation. We have implemented the outlined strategy with the setup as follows.

- the sizes of the sensing matrices A were $(m = 96) \times (n = 128)$, with $B = I$ and $K = 32$ four-element blocks in $Bx = x$;
- the 96×128 sensing matrices A were built as follows: we first draw a matrix at random from one of the following distributions:
 - type H: randomly selected 96×128 submatrix of the 128×128 Hadamard matrix⁵,
 - type G: 96×128 matrix with independent $\mathcal{N}(0, 1)$ entries,
 - type R: 96×128 matrix with independent entries taking values ± 1 with equal probabilities,
 - type T: random 96×128 matrix of the structure arising in *Multi-Task Learning* (see, e.g., [1] and references therein): the consecutive 4-column parts of the matrix are block-diagonal with four 24×1 diagonal blocks with independent $\mathcal{N}(0, 1)$ entries,

and then scale the columns of the selected matrix to have their $\|\cdot\|_2$ -norms equal to 1;

The results we report describe 4 experiments differing from each other by the type of the (randomly selected) matrix A .⁶

⁵The Hadamard matrices H_k of order $2^k \times 2^k$, $k = 0, 1, \dots$, are given by the recurrence $H_0 = 1$, $H_{k+1} = [H_k, H_k; H_k, -H_k]$. They are symmetric matrices with ± 1 entries and rows orthogonal to each other.

⁶As far as our experience shows, the results remain nearly the same across instances of A drawn from the same distribution, so that only one experiment for each type of distributions in question appears to be representative enough.

A	r	$H^{(\text{MI})}$			$H^{(\text{MBI})}$			$H^{(1)}$			$H^{(2)}$			$H^{(\infty)}$			$\bar{s}(r)$
H	1	2	0.4727	0.509	2	0.444	0.460	<u>3</u>	0.429	0.429	2	0.487	0.519	<u>3</u>	0.429	0.429	4
	2	2	0.436	0.436	2	0.429	0.429	<u>3</u>	0.429	0.429	<u>3</u>	0.429	0.429	<u>3</u>	0.429	0.429	3
	∞	2	0.473	0.509	2	0.444	0.460	<u>3</u>	0.429	0.429	2	0.487	0.519	<u>3</u>	0.429	0.429	3
G	1	0	0.000	0.000	0	0.000	0.000	<u>3</u>	0.467	0.900	1	0.301	0.301	1	0.489	0.489	5
	2	0	0.000	0.000	1	0.368	0.368	1	0.300	0.300	<u>3</u>	0.447	0.458	2	0.479	0.549	5
	∞	0	0.000	0.000	0	0.000	0.000	0	0.000	0.000	1	0.305	0.305	<u>3</u>	0.483	0.823	4
R	1	0	0.0000	0.000	0	0.000	0.000	<u>3</u>	0.477	0.853	1	0.291	0.291	1	0.498	0.498	5
	2	0	0.000	0.000	1	0.354	0.354	1	0.284	0.284	<u>3</u>	0.438	0.440	1	0.264	0.264	5
	∞	0	0.000	0.000	0	0.000	0.000	1	0.482	0.482	1	0.286	0.286	<u>3</u>	0.489	0.739	5
T	1	1	0.384	0.384	1	0.399	0.399	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	3
	2	1	0.384	0.384	1	0.399	0.399	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	3
	∞	1	0.384	0.384	1	0.399	0.399	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	<u>2</u>	0.383	0.383	3

Table 1: Certified sparsity levels for penalized ℓ_1/ℓ_r -recoveries for candidate contrast matrices. For each candidate and each value of r we present in the corresponding cells the triple $s(H, r) | \kappa_1^{r, s(H, r)}(H) | \kappa_\infty^{r, s(H, r)}(H)$. $\bar{s}(r)$: a computed upper bound on r -goodness $s^*(A, r)$ of A . Underlined in red: the best sparsity $s_*(\mathcal{H})$ certified by our sufficient conditions for the validity of penalized recovery.

In Table 1, we display the certified sparsity levels of penalized ℓ_1/ℓ_r recoveries for the candidate contrast matrices. In addition, we present valid *upper bounds* $\bar{s}(r)$ on the “ r -goodness” $s^*(A, r)$ of A , defined as the largest s such that the ℓ_1/ℓ_r -recovery in the noiseless case recovers *exactly* the representations of *all* s -block-sparse vectors, that is,

$$s^*(A, r) = \max \left\{ s : x = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \sum_{k=1}^K \|[z]_k\|_r : Az = Ax \right\} \text{ for all } s\text{-block-sparse } x. \right\}$$

We present on Figure 1 examples of “bad” signals (i.e., $(\bar{s}(r) + 1)$ -block-sparse signals which are *not* recovered correctly by the latter procedure).⁷

On the basis of this experiment we can make two tentative conclusions:

- the ℓ_1/ℓ_2 recovery with the contrast matrix $H^{(2)}$ and the ℓ_1/ℓ_∞ recovery with the contrast matrix $H^{(\infty)}$ were able to certify the best levels of allowed sparsity (when compared to other candidate matrices from \mathcal{H});
- in our experiments, the upper bounds $\bar{s}(r)$ on the r -goodness $s^*(A, r)$ of A are close to the corresponding certified lower bounds $s_*(\mathcal{H}, r) = \max_{H \in \mathcal{H}} s(H, r)$.

Numerical evaluation of recovery errors. The objective of the next experiment is to evaluate the accuracy of penalized ℓ_1/ℓ_r recoveries in the noisy setting. We consider the contrast matrices from $\mathcal{H} = \{H^{(\text{MI})}, H^{(\text{MBI})}, H^{(1)}, H^{(2)}, H^{(\infty)}\}$ as above. Note that it is possible to improve the error bound by optimizing it over H as it was done in section 5.2. In the experiments to be reported this additional optimization, however, did not yield a significant improvement (which perhaps reflects the “nice conditioning” of the sensing matrices we dealt with), and we do not present the simulation results for optimized contrasts here.

- We ran four series of simulations corresponding to the four instances of the sensing matrix A we used. The series associated with a particular A was as follows:

⁷It is immediately seen that whenever B is of full row rank, the *nullspace property* “ $L_{s,1}(Bx) < \frac{1}{2}L_1(Bx)$ for all $x \in \operatorname{Ker} A$ with $Bx \neq 0$ ” is necessary for s to be $\leq s^*(A, \cdot)$. As a result, for B ’s of full row rank, $s^*(A, r)$ can be upper-bounded in a manner completely similar to the case of the standard r.s., see [23, section 4.1]

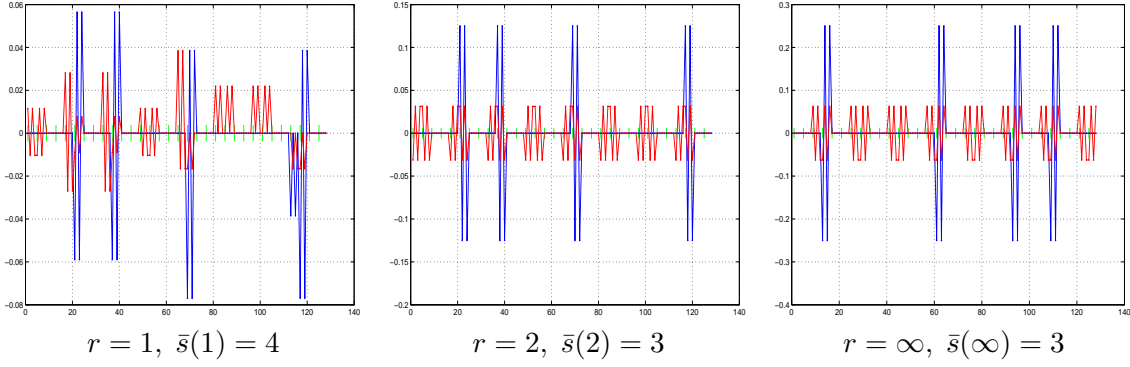


Figure 1: "Bad" $(\bar{s}(r) + 1)$ -block-sparse signals (blue) and their ℓ_1/ℓ_r recoveries (red) from noiseless observations, H -matrix A .

- Given A , we associate with it the five aforementioned candidate contrast matrices from \mathcal{H} . Combining these matrices with 3 values of r ($r = 1, 2, \infty$), we get 15 recovery routines. We augmented these 15 routines by the block Lasso recovery as described in [26]. In our notation, this recovery is (cf. [26, (2.2)])

$$\hat{x}_{\text{Lasso}}(y) \in \underset{z}{\text{Argmin}} \left\{ \frac{1}{m} \|Az - y\|_2^2 + 2 \sum_{k=1}^K \lambda_k \|z[k]\|_2 \right\} \quad (7.50)$$

($z[k]$, $1 \leq k \leq K$, are the blocks in $z = B \textcolor{blue}{z} \textcolor{blue}{x}$), with the penalty coefficients λ_k chosen according to the equality version of the relations in [26, Theorem 3.1] used with $q = 2$.

Each of the 16 resulting recovery routines was on two samples, each of size 100, of randomly generated recovery problems as follows. In each problem the true signal was randomly generated with s nonzero blocks, and the observations were affected by pure Gaussian white noise: $y = Ax + \sigma\xi$, $\xi \sim \mathcal{N}(0, I)$. In the first sample, s was set to the best value $s_*(\mathcal{H})$ of block sparsity we were able to certify; in the second, $s = 2s_*(\mathcal{H})$ was used. The parameter λ in the penalized recoveries was set to $2s$ (and thus was tuned to the actual sparsity of test signals). In both samples, we used $\sigma = 0.001$.

We compare the recovery routines on the basis of their *ratings* computed as follows: given a recovery problem from the sample, we applied to it every one of our 16 recovery routines and measured the 16 resulting $\|\cdot\|_\infty$ -errors. Dividing the smallest of these errors by the error of a given routine we obtain "the rating" of the routine in this particular simulation. Thus, all ratings are ≤ 1 ; the routine which attains the best $\|\cdot\|_\infty$ -recovery error for the current data is rated "1.0". For the remaining routines, the closer to 1 is the rating of the routine, the closer is the routine to the "winner" of the current simulation. The final rating of a given recovery routine is its average rating over all $800 = 4 \times 2 \times 100$ recovery problems processed in the experiment.

The resulting ratings are presented in Table 2. The "winner" is the routine associated with $r = 2$ and $H = H^{(2)}$. Surprisingly, the second best routine is associated with the same $r = 2$ and the simplest contrast $H^{(\text{MI})}$, an outsider in terms of the data presented in Table 1. This inconsistency may be explain by the fact that the data in Table 1 describe the guaranteed worst-case behavior of our recovery routines, which may be quite different from their "average behavior", reflected by Table 2. Our tentative conclusion on the basis of the data from Tables 1, 2 is that the penalized ℓ_1/ℓ_2 recovery associated with the contrast matrix $H^{(2)}$ may be favored when recovery guarantees are to be associated with good numerical performance.

The above comparison was carried out for σ set to 0.001. The conducted experiments show that for the routines in question and our purely Gaussian model of observation errors, the recovery errors are, typically, proportional to σ . This is illustrated by the plots on Figure 2 on which we traced the average (over 40 experiments for every grid value of σ) *signal-to-noise ratio* (the ratio of the $\|\cdot\|_\infty$ -error of the recovery to σ) of our favorable recovery ($r = 2$, $H = H^{(2)}$) and the corresponding performance figure for block Lasso.

r	$H^{(\text{MI})}$	$H^{(\text{MBI})}$	$H^{(1)}$	$H^{(2)}$	$H^{(\infty)}$	Lasso
1	0.30	0.20	0.53	0.60	0.54	N/A
2	0.76	0.51	0.75	0.79	0.75	0.19
∞	0.25	0.18	0.44	0.48	0.44	N/A

Table 2: Ratings of recovery routines.

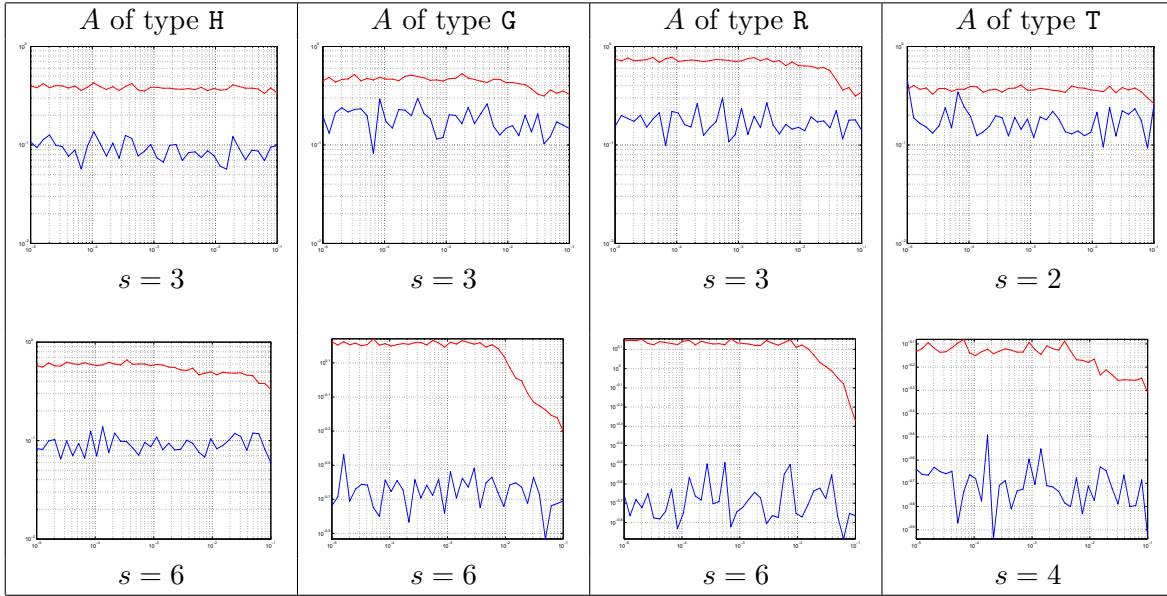


Figure 2: Average over 40 experiments ratio of $\|\cdot\|_\infty$ -recovery error to σ vs. σ . In blue: ℓ_1/ℓ_2 recovery, $H = H^{(2)}$; in red: Lasso recovery.

References

- [1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.
- [2] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.
- [3] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982–2001, 2010.
- [4] Z. Ben-Haim and Y. C. Eldar. Near-oracle performance of greedy block-sparse estimation techniques from noisy measurements. Technical report, 2010. <http://arxiv.org/abs/1009.0906>.
- [5] A. Ben-Tal, L. E. Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.
- [6] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Stat.*, 37(4):1705–1732, 2008.
- [7] P. Bühlmann and S. van de Geer. On the conditions used to prove oracle results for the Lasso. *Electron. J. Statist.*, 3:1360–1392, 2009.
- [8] E. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(6):4203–4215, 2005.
- [9] E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n . *Annals of Stat.*, 35(6):2313–2351, 2007.
- [10] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus de l’Acad. des Sci., Serie I(346)*:589–592, 2008.
- [11] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51:4203–4215, 2006.
- [12] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables Lasso. *Mathematical Methods of Statistics*, 27(4):317–326, 2008.
- [13] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, 52:6–18, 2006.
- [14] M. Duarte, W. Bajwa, and R. Calderbank. The performance of group Lasso for linear regression of grouped variables. Technical report, 2011. TR-2010-10, Department of Computer Science, Duke University, Durham, NC, http://www.duke.edu/~wb40/pubs/sampta11_tr.pdf.
- [15] Y. C. Eldar, P. Kuppinger, and H. Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. on Signal Processing*, 58(6):3042–3054, 2010.
- [16] Y. C. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Trans. Inf. Theory*, 55(11):5302–5316, 2009.
- [17] E. Elhamifar and R. Vidal. Structured sparse recovery via convex optimization. Technical report, 2011. <http://arxiv.org/pdf/1104.0654>.
- [18] R. Gribonval and R. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inf. Theory*, 49:3320–3325, 2003.
- [19] J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Stat.*, 38(4):1978–2004, 2010.

- [20] G. M. James, P. Radchenko, and J. Lv. Dasso: connections between the Dantzig selector and Lasso. *J. Roy. Statist. Soc. Ser. B*, 71(1):127–142, 2009.
- [21] A. B. Juditsky, F. Kılınç-Karzan, and A. S. Nemirovski. Verifiable conditions of ℓ_1 recovery for sparse signals with sign restrictions. *Math. Progr.*, 127(1):89–122, 2010. http://www.optimization-online.org/DB_HTML/2009/03/2272.html.
- [22] A. B. Juditsky and A. S. Nemirovski. Accuracy guarantees for ℓ_1 -recovery. Technical report, 2010. http://www.optimization-online.org/DB_HTML/2010/10/2778.html.
- [23] A. B. Juditsky and A. S. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Progr.*, 127(1):57–88, 2010. Special issue on machine learning.
- [24] H. Liu and J. Zhang. Estimation consistency of the group Lasso and its applications. *Journal of Machine Learning Research - Proceedings Track*, 5:376–383, 2009.
- [25] H. Liu, J. Zhang, X. Jiang, and J. Liu. The group Dantzig selector. *Journal of Machine Learning Research - Proceedings Track*, 9:461–468, 2010.
- [26] K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer. Oracle inequalities and optimal inference under group sparsity. Technical report, 2010. <http://arxiv.org/pdf/1007.1771>.
- [27] L. Meier, S. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B, Methodological*, 70(1):53–71, 2008.
- [28] Y. Nardi and A. Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electron. J. Statist.*, 2:605–633, 2008.
- [29] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Stat.*, 39(1):1–47, 2011.
- [30] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi. Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays. *IEEE J. Selected Topics in Signal Processing*, 2(3):275–285, 2008.
- [31] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. on Signal Processing*, 57(8):3075–3085, 2009.
- [32] H. Wolkowicz, R. Saigal, and E. Lieven Vandenberghe. *Handbook on semidefinite programming*. Kluwer Academic Publishers, 2000.
- [33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. Roy. Stat. Soc. Ser. B*, 68(1):49–67, 2006.

A Proofs

A.1 Proof of Proposition 2.1

Let $x \in \mathbb{R}^n$, and let x^1, \dots, x^q be obtained from x by the following construction: x^1 is obtained from x by zeroing all but the s largest in magnitude blocks; x^2 is obtained by the same procedure applied to $x - x^1$, x^3 – by the same procedure applied to $x - x^1 - x^2$, and so on; the process is terminated at the first step q when it happens that $x = x^1 + \dots + x^q$. Note for $j \geq 2$ we have $L_\infty(x^j) \leq s^{-1}L_1(x^{j-1})$ and $L_1(x^j) \leq L_1(x^{j-1})$,

whence also $\|x^j\|_2 = L_2(x^j) \leq \sqrt{L_\infty(x^j)L_1(x^j)} \leq s^{-1/2}L_1(x^{j-1})$. Recall that if A is BRIP($\delta, 2s$), then for every two s -block-sparse vectors u, v with non-overlapping supports we have

$$|u^T A^T A u| \leq \delta \|u\|_2 \|v\|_2. \quad (*)$$

(i): We have

$$\begin{aligned} & \|Ax^1\|_2 \|Ax\|_2 \geq [x^1]^T A^T A x = \|Ax^1\|_2^2 - \sum_{j=2}^q [x^1]^T A^T A x^j \\ & \geq \|Ax^1\|_2^2 - \delta \sum_{j=2}^t \|x^1\|_2 \|x^j\|_2 \text{ [by (*)]} \\ & \geq \|Ax^1\|_2^2 - \delta s^{-1/2} \|x^1\|_2 \sum_{j=2}^q L_1(x^{j-1}) \geq \|Ax^1\|_2^2 - \delta s^{-1/2} \|x^1\|_2 L_1(x) \\ \Rightarrow & \|Ax^1\|_2^2 \leq \|Ax^1\|_2 \|Ax\|_2 + \delta s^{-1/2} \|x^1\|_2 L_1(x) \\ \Rightarrow & \|x^1\|_2 = \frac{\|x^1\|_2}{\|Ax^1\|_2} \|Ax^1\|_2^2 \leq \frac{\|x^1\|_2}{\|Ax^1\|_2} \|Ax\|_2 + \delta s^{-1/2} \left(\frac{\|x^1\|_2}{\|Ax^1\|_2} \right)^2 L_1(x) \\ \Rightarrow & L_{s,2}(x) = \|x^1\|_2 \leq \frac{1}{\sqrt{1-\delta}} \|Ax\|_2 + \frac{\delta s^{-1/2}}{1-\delta} L_1(x) \text{ [by BRIP}(\delta, 2s)\text{]} \end{aligned}$$

and we see that the pair $\left(H = \frac{s^{-1/2}}{\sqrt{1-\delta}} I_m, \|\cdot\|_2\right)$ satisfies $\mathbf{Q}_{s,2}(\frac{\delta}{1-\delta})$, as claimed in (i).

(ii): We have

$$\begin{aligned} & L_1(x^1) L_\infty(A^T A x) \geq [x^1]^T A^T A x = \|Ax^1\|_2^2 - \sum_{j=2}^q [x^1]^T A^T A x^j \\ & \geq \|Ax^1\|_2^2 - \delta s^{-1/2} \|x^1\|_2 L_1(x) \text{ [exactly as above]} \\ \Rightarrow & \|Ax^1\|_2^2 \leq L_1(x^1) L_\infty(A^T A x) + \delta s^{-1/2} \|x^1\|_2 L_1(x) \\ \Rightarrow & (1-\delta) \|x^1\|_2^2 \leq L_1(x^1) L_\infty(A^T A x) + \delta s^{-1/2} \|x^1\|_2 L_1(x) \text{ [by BRIP}(\delta, 2s)\text{]} \\ & \leq s^{1/2} \|x^1\|_2 L_\infty(A^T A x) + \delta s^{-1/2} \|x^1\|_2 L_1(x) \\ \Rightarrow & L_{s,2}(x) = \|x^1\|_2 \leq \frac{s^{1/2}}{1-\delta} L_\infty(A^T A x) + \frac{\delta}{1-\delta} s^{-1/2} L_1(x), \end{aligned}$$

and we see that the pair $\left(H = \frac{1}{1-\delta} A, L_\infty(\cdot)\right)$ satisfies the condition $\mathbf{Q}_{s,2}(\frac{\delta}{1-\delta})$, as required in (ii). \blacksquare

A.2 Proof of Theorems 3.1 and 3.2

All we need is to prove Theorem 3.2, since Theorem 3.1 is the particular case $\varkappa = \kappa < 1/2$ of Theorem 3.2.

Let us fix $x \in \mathbb{R}^n$, $u \in \mathcal{U}$ and $\xi \in \Xi$, and let us set $\eta = u + \sigma \xi$, $\hat{x} = \hat{x}_{\text{reg}}(Ax + \eta)$. Let also $I \in \{1, \dots, K\}$ be the set of indexes of the s largest in magnitude blocks in Bx , J be the complement of I in $\{1, \dots, K\}$, and let for $w \in \mathcal{W}$, w_I and w_J be the vectors obtained from w by zeroing blocks $w[k]$ with indices $k \notin I$ and $k \notin J$, respectively, and keeping the remaining blocks intact. Finally, let $z = \hat{x} - x$.

1⁰. By the definition of Ξ and due to $\xi \in \Xi$, $u \in \mathcal{U}$, we have

$$\|H^T([Ax + \eta] - Ax)\| \leq \rho, \quad (\text{A.51})$$

so that x is a feasible solution to the optimization problem specifying \hat{x} , whence $L_1(B\hat{x}) \leq L_1(Bx)$. We therefore have

$$\begin{aligned} L_1([B\hat{x}]_J) &= L_1(B\hat{x}) - L_1([B\hat{x}]_I) \leq L_1(Bx) - L_1([B\hat{x}]_I) \\ &= L_1([Bx]_I) + L_1([Bx]_J) - L_1([B\hat{x}]_I) \leq L_1([Bz]_I) + L_1([Bx]_J), \end{aligned}$$

and therefore

$$L_1([Bz]_J) \leq L_1([B\hat{x}]_J) + L_1([Bx]_J) \leq L_1([Bz]_I) + 2L_1([Bx]_J).$$

It follows that

$$L_1(Bz) = L_1([Bz]_I) + L_1([Bz]_J) \leq 2L_1([Bz]_I) + 2L_1([Bx]_J). \quad (\text{A.52})$$

Further, by definition of \hat{x} we have $\|H^T([Ax + u + \xi] - A\hat{x})\| \leq \rho$, which combines with (A.51) to imply that

$$\|H^T A(\hat{x} - x)\| \leq 2\rho. \quad (\text{A.53})$$

2⁰. Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,1}(\varkappa)$, we have

$$L_{s,1}(Bz) \leq s\|H^T Az\| + \varkappa L_1(Bz).$$

By (A.53), it follows that $L_{s,1}(Bz) \leq 2s\rho + \varkappa L_1(Bz)$, which combines with the evident inequality $L_1([Bz]_I) \leq L_{s,1}(Bz)$ and with (A.52) to imply that

$$L_1([Bz]_I) \leq 2s\rho + \varkappa L_1(Bz) \leq 2s\rho + 2\varkappa L_1([Bz]_I) + 2\varkappa L_1([Bx]_J),$$

whence

$$L_1([Bz]_I) \leq \frac{2s\rho + 2\varkappa L_1([Bx]_J)}{1 - 2\varkappa}.$$

Invoking (A.52), we conclude that

$$L_1(Bz) \leq \frac{4s}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right]. \quad (\text{A.54})$$

3⁰. Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,q}(\kappa)$, we have

$$L_{s,q}(Bz) \leq s^{\frac{1}{q}} \|H^T Az\| + \kappa s^{\frac{1}{q}-1} L_1(Bz),$$

which combines with (A.54) and (A.53) to imply that

$$L_{s,q}(Bz) \leq s^{\frac{1}{q}} 2\rho + \kappa s^{\frac{1}{q}} \frac{4\rho + 2s^{-1} L_1([Bx]_J)}{1 - 2\varkappa} \leq \frac{4s^{\frac{1}{q}} [1 + \kappa - \varkappa]}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right] \quad (\text{A.55})$$

(we have taken into account that $\varkappa < 1/2$ and $\kappa \geq \varkappa$). Let θ be the $(s+1)$ -st largest magnitude of the blocks of Bz , and let $w = Bz - [Bz]^s$. Now (A.55) implies that

$$\theta \leq L_{s,q}(Bz) s^{-\frac{1}{q}} \leq \frac{4[1 + \kappa - \varkappa]}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right].$$

Hence we have

$$\begin{aligned} L_q(w) &\leq L_\infty(w)^{\frac{q-1}{q}} L_1(w)^{\frac{1}{q}} \leq \theta^{\frac{q-1}{q}} L_1(Bz)^{\frac{1}{q}} \leq \theta^{\frac{q-1}{q}} \frac{(4s)^{\frac{1}{q}}}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right]^{\frac{1}{q}} \\ &\leq \frac{4s^{\frac{1}{q}} [1 + \kappa - \varkappa]^{\frac{q-1}{q}}}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right]. \end{aligned}$$

Taking into account (A.55) and the fact that the supports of $[Bz]^s$ and w do not intersect, we get

$$\begin{aligned} L_q(Bz) &\leq 2^{\frac{1}{q}} \max[L_q([Bz]^s), L_q(w)] = 2^{\frac{1}{q}} \max[L_{s,q}(Bz), L_q(w)] \\ &\leq \frac{4(2s)^{\frac{1}{q}} [1 + \kappa - \varkappa]}{1 - 2\varkappa} \left[\rho + \frac{1}{2s} L_1([Bx]_J) \right]. \end{aligned}$$

This bound combines with (A.54), the Hölder inequality and the relation $L_1([Bx]_J) = L_1(Bx - [Bx]^s)$ to imply (3.13). ■

A.3 Proof of Theorem 3.3

Let us prove (i). Let us fix $x \in \mathbb{R}^n$, $u \in \mathcal{U}$ and ξ , and let us set $\eta = u + \xi$, $\hat{x} = \hat{x}_{\text{pen}}(Ax + \eta)$. Let also $I \subset \{1, \dots, K\}$ be the set of indices of the s largest in magnitude blocks in Bx , J be the complement of I in $\{1, \dots, K\}$, and for $w \in \mathcal{W}$ let w_I , w_J be the vectors obtained from w by zeroing out all blocks with indexes not in I , respectively, not in J . Finally, let $z = \hat{x} - x$ and $\nu = \|H^T \eta\|$.

1^0 . We have

$$L_1(B\hat{x}) + \lambda \|H^T(A\hat{x} - Ax - \eta)\| \leq L_1(Bx) + \lambda \|H^T \eta\|$$

and

$$\|H^T(A\hat{x} - Ax - \eta)\| = \|H^T(Az - \eta)\| \geq \|H^T Az\| - \|H^T \eta\|,$$

whence

$$L_1(B\hat{x}) + \lambda \|H^T Az\| \leq L_1(Bx) + 2\lambda \|H^T \eta\| = L_1(Bx) + 2\lambda \nu. \quad (\text{A.56})$$

We have

$$\begin{aligned} L_1(B\hat{x}) &= L_1(Bx + Bz) = L_1([Bx]_I + [Bz]_I) + L_1([Bx]_J + [Bz]_J) \\ &\geq L_1([Bx]_I) - L_1([Bz]_I) + L_1([Bz]_J) - L_1([Bx]_J), \end{aligned}$$

which combines with (A.56) to imply that

$$L_1([Bx]_I) - L_1([Bz]_I) + L_1([Bz]_J) - L_1([Bx]_J) + \lambda \|H^T Az\| \leq L_1(Bx) + 2\lambda \nu,$$

or, which is the same,

$$L_1([Bz]_J) - L_1([Bz]_I) + \lambda \|H^T Az\| \leq 2L_1([Bx]_J) + 2\lambda \nu. \quad (\text{A.57})$$

Since $(H, \|\cdot\|)$ satisfies $\mathbf{Q}_{s,1}(\varkappa)$, we have

$$L_1([Bz]_I) \leq L_{s,1}(Bz) \leq s \|H^T Az\| + \varkappa L_1(Bz),$$

so that

$$(1 - \varkappa)L_1([Bz]_I) - \varkappa L_1([Bz]_J) - s \|H^T Az\| \leq 0. \quad (\text{A.58})$$

Taking weighted sum of (A.57) and (A.58), the weights being 1, 2, respectively, we get

$$(1 - 2\varkappa)[L_1([Bz]_I) + L_1([Bz]_J)] + (\lambda - 2s)\|H^T Az\| \leq 2L_1([Bx]_J) + 2\lambda \nu,$$

that is (since $\lambda \geq 2s$),

$$L_1(Bz) \leq \frac{2\lambda \nu + 2L_1([Bx]_J)}{1 - 2\varkappa} = \frac{2\lambda}{1 - 2\varkappa} \left[\nu + \frac{1}{2s} L_1([Bx]_J) \right]. \quad (\text{A.59})$$

Further, by (A.56) we have

$$\lambda \|H^T Az\| \leq L_1(Bx) - L_1(B\hat{x}) + 2\lambda \nu \leq L_1(Bz) + 2\lambda \nu,$$

which combines with (A.59) to imply that

$$\lambda \|H^T Az\| \leq \frac{2\lambda \nu + 2L_1([Bx]_J)}{1 - 2\varkappa} + 2\lambda \nu = \frac{2\lambda \nu(2 - 2\varkappa) + 2L_1([Bx]_J)}{1 - 2\varkappa}. \quad (\text{A.60})$$

From $\mathbf{Q}_{s,q}(\kappa)$ it follows that

$$L_{s,q}(Bz) \leq s^{\frac{1}{q}} \|H^T Az\| + \kappa s^{\frac{1}{q}-1} L_1(Bz),$$

which combines with (A.60) and (A.59) to imply that

$$\begin{aligned}
L_{s,q}(Bz) &\leq s^{\frac{1}{q}-1} [s\|H^T Ax\| + \kappa L_1(Bz)] \leq s^{\frac{1}{q}-1} \left[\frac{4s\nu(1-\kappa) + \frac{2s}{\lambda} L_1([Bx]_J)}{1-2\kappa} + \frac{\kappa[2\lambda\nu + 2L_1([Bx]_J)]}{1-2\kappa} \right] \\
&= \frac{s^{\frac{1}{q}} [4\nu(1-\kappa) + 2s^{-1}\lambda\kappa\nu] + 2[\lambda^{-1} + s^{-1}\kappa] L_1([Bx]_J)}{1-2\kappa} \\
&\leq 4 \frac{s^{\frac{1}{q}}}{1-2\kappa} \left[1 + \frac{\kappa\lambda}{2s} - \kappa \right] \left[\nu + \frac{1}{2s} L_1([Bx]_J) \right] \tag{A.61}
\end{aligned}$$

(recall that $\lambda \geq 2s$, $\kappa \geq \kappa$, and $\kappa < 1/2$). It remains to repeat the reasoning following (A.55) in item 3⁰ of the proof of Theorem 3.2. Specifically, denoting by θ the $(s+1)$ -st largest magnitude of blocks of Bz , (A.61) implies that

$$\theta \leq s^{-1/q} L_{s,q}(Bz) \leq 4[1 + \kappa \frac{\lambda}{2s} - \kappa] [\nu + \frac{1}{2s} L_1([Bx]_J)], \tag{A.62}$$

so that for the vector $w = Bz - [Bz]^s$ one has

$$L_q(w) \leq \theta^{1-\frac{1}{q}} L_1(w)^{\frac{1}{q}} \leq \frac{4(\lambda/2)^{\frac{1}{q}}}{1-2\kappa} [1 + \kappa \frac{\lambda}{2s} - \kappa]^{\frac{q-1}{q}} [\nu + \frac{1}{2s} L_1([Bx]_J)]$$

(we have used (A.62), (A.59) and the fact that $\lambda \geq 2s$). Hence, taking into account that $[Bz]^s$ and w have non-intersecting supports,

$$\begin{aligned}
L_q(Bz) &\leq 2^{\frac{1}{q}} \max[L_q([Bz]^s), L_q(w)] = 2^{\frac{1}{q}} \max[L_{s,q}(Bz), L_q(w)] \\
&\leq \frac{4\lambda^{\frac{1}{q}}}{1-2\kappa} \left[1 + \kappa \frac{\lambda}{2s} - \kappa \right] \left[\nu + \frac{1}{2s} L_1([Bx]_J) \right]
\end{aligned}$$

(we have used (A.61)). This combines with (A.59) and Hölder inequality to imply (3.16). All remaining claims of Theorem 3.3 are immediate corollaries of (3.16). \blacksquare

A.4 Proof of Proposition 4.1

(i): Let $H \in \mathbb{R}^{m \times M}$, $\|\cdot\|$, ρ satisfy (!). Then for every $k \leq K$ and every $1 \leq i \leq n_i$, denoting by w_{ki} i -th entry in $w[k]$, $w \in \mathcal{W}$, we have

$$|[Bx]_{ki}| \leq \|H^T Ax\| + s^{-1}\kappa L_1(Bx),$$

or, which is the same by homogeneity,

$$\min_x \{ \|H^T Ax\| - [Bx]_{ki} : L_1(Bx) \leq 1 \} \geq -s^{-1}\kappa.$$

In other words, the optimal value Opt_{ki} of the conic optimization problem

$$\text{Opt}_{ki} = \min_{x,t} \left\{ t - [e^{ki}]^T Bx : \|H^T Ax\| \leq t, L_1(Bx) \leq 1 \right\},$$

where $e^{ki} \in \mathcal{W}$ is the vector with the only nonzero entry, equal to 1, placed at i -th position of the k -th block, is $\geq -s^{-1}\kappa$. Since the problem clearly is strictly feasible, this is the same as to say that the dual problem

$$\max_{\mu \in \mathbb{R}, g \in \mathcal{W}, \eta \in \mathbb{R}^M} \left\{ -\mu : A^T H \eta + B^T g = B^T e^{ki}, \|g[\ell]\|_1 \leq \mu, 1 \leq \ell \leq K, \|\eta\|_* \leq 1 \right\},$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$, has a feasible solution with the value of the objective $\geq -s^{-1}\kappa$. It follows that there exists $\eta = \eta^{ki}$ and $g = g^{ki}$ such that

$$(a) : B^T e^{ki} = A^T h^{ki} + B^T g^{ki}, \quad (b) : h^{ki} := H \eta^{ki}, \|\eta^{ki}\|_* \leq 1, \quad (c) : \|g^{ki}[\ell]\|_1 \leq s^{-1}\kappa, 1 \leq \ell \leq K. \tag{A.63}$$

Denoting by h^i the i -th column in the $m \times N$ -matrix $[h^{1,1}, \dots, h^{1,n_1}, h^{2,1}, \dots, h^{2,n_2}, \dots, h^{K,1}, \dots, h^{K,n_K}]$, defining $V^{k\ell}$ as the $n_k \times n_\ell$ matrix with the rows $(g^{ki}[\ell])^T$, $i = 1, \dots, n_k$, and setting $V = [V^{k\ell}]_{k,\ell=1}^K$, (A.63.a,c) ensure the validity of (4.20a,b) (recall that $\|M\|_{\infty,\infty}$ is the maximum of $\|\cdot\|_1$ -norms of the rows in M). Besides, by (A.63.b) and the definition of Ξ (see (!)) we have

$$\xi \in \Xi \Rightarrow \|H^T[u + \xi]\| \leq \rho \ \forall u \in \mathcal{U} \Rightarrow_a \|[h^{ki}]^T[u + \xi]\| \leq \rho \ \forall u \in \mathcal{U} \Rightarrow_b \max_{u \in \mathcal{U}} u^T h^{ki} + |\xi^T h^{ki}| \leq \rho$$

where the implication \Rightarrow_a is due to the fact that $\|[h^{ki}]^T \zeta\| = |[\eta^{ki}]^T H^T \zeta| \leq \|H^T \zeta\|$ for all ζ because of $\|\eta^{ki}\|_* \leq 1$, and the implication \Rightarrow_b is due to the fact that \mathcal{U} is symmetric w.r.t. the origin. We conclude that $\Xi \subset \Xi^+$ and thus $P(\Xi^+) \geq P(\Xi) \geq 1 - \epsilon$, as required in (4.20.c). (i) is proved.

(ii): Let $\hat{H} = [h^1, \dots, h^N]$, $V = [V^{k\ell}]_{k,\ell=1}^K$, ρ satisfy (4.20). Then for every $x \in \mathbb{R}^n$ we have

$$w := Bx = VBx + \hat{H}^T Ax = Vw + \underbrace{\hat{H}^T Ax}_v,$$

whence $w[k] = \sum_{\ell=1}^K V^{k\ell} w[\ell] + v[k]$, so that

$$\|w[k]\|_{(k)} = \|w[k]\|_\infty \leq \sum_{\ell=1}^K \|V^{k\ell}\|_{\infty,\infty} \|w[\ell]\|_\infty + \|v[k]\|_\infty \leq s^{-1} \kappa L_1(w) + \|\hat{H}^T Ax\|_\infty.$$

We conclude that

$$L_{s,\infty}(Bx) = \max_k \|w[k]\|_{(k)} \leq \|\hat{H}^T Ax\|_\infty + s^{-1} \kappa L_1(Bx)$$

for all x , meaning that $(\hat{H}, \|\cdot\|_\infty)$ satisfies $\mathbf{Q}_{s,\infty}(\kappa)$. Further, we have

$$\begin{aligned} \Xi &:= \left\{ \xi : \|\hat{H}^T[u + \xi]\|_\infty \leq \rho \ \forall u \in \mathcal{U} \right\} = \left\{ \xi : |[h^i]^T[u + \xi]| \leq \rho \ \forall u \in \mathcal{U}, \ \forall i \leq N \right\} \\ &= \left\{ \xi : \max_{u \in \mathcal{U}} u^T h^i + |[h^i]^T \xi| \leq \rho \ \forall i \leq N \right\} = \Xi^+, \end{aligned}$$

whence $P(\Xi) = P(\Xi^+) \geq 1 - \epsilon$. Thus, $\hat{H}, \|\cdot\|_\infty, \rho$ satisfy (!). ■

A.5 Proof of Proposition 4.2

Notation. Let $1 \leq k \leq K$ and $1 \leq i \leq n_k$. For a vector $w \in \mathcal{W}$, we set $[w]_{ki}$ to be i -th coordinate in $w[k]$. For a vector $u \in \mathbb{R}^{n_k}$, we set $\|u\|_\infty^i = \max_{j \neq i} |u_j|$, with the convention that the latter maximum is 0 when $n_k = 1$. Further, let e^{ki} be the vector from \mathcal{W} such that $[e^{ki}]_{\ell j} = 1$ when $\ell = k$ and $j = i$ and $[e^{ki}]_{\ell j} = 0$ for all remaining pairs ℓ, j . Finally, let $B = [B^1; \dots; B^K]$ with $n_k \times n$ matrices B^k .

1⁰. Let us fix k, i , $1 \leq k \leq K$, $1 \leq i \leq n_k$, let $M = \alpha S$, and let $a > \alpha + S^{-1}M = 2\alpha$. We set

$$\begin{aligned} X_+^{ki} &= \{x \in \mathbb{R}^n : [B^k x]_i = a, \|B^k x\|_\infty^i + \sum_{\ell \neq k} \|B^\ell x\|_\infty \leq M\}, \quad X_-^{ki} = -X_+^{ki}, \\ Y_+^{ki} &= AX_+^{ki}, \quad Y_-^{ki} = AX_-^{ki} = -Y_+^{ki}. \end{aligned}$$

We denote $\mathcal{V} = 2\mathcal{U} + 2\{D\eta : \|\eta\|_2 \leq \text{Erfinv}(\epsilon)\}$.

It may happen that $X_\pm^{ki} = \emptyset$. This is exactly the same as to say that the optimal value in the strictly feasible conic optimization problem

$$\max_x \left\{ [e^{ki}]^T Bx : \|B^k x\|_\infty^i + \sum_{\ell \neq k} \|B^\ell x\|_\infty \leq M \right\}$$

is $< a$, meaning that the optimal value in the dual problem

$$\min_{v \in \mathcal{W}, t} \left\{ Mt : [v]_{ki} = 0, \sum_{\ell=1}^K [B^\ell]^T v[\ell] = B^T e^{ki}, \max_{1 \leq \ell \leq K} \|v[\ell]\|_1 \leq t \right\}$$

is $< a$, whence there exists $v^{ki} \in \mathcal{W}$ such that $[v^{ki}]_{ki} = 0$, $B^T v^{ki} = B^T e^{ki}$ and $M \max_{\ell} \|v^{ki}[\ell]\|_1 < a$, that is, $\max_{\ell} \|v^{ki}[\ell]\|_1 < a/M$. Thus, when X_{\pm}^{ki} is empty, setting $h^{ki} = 0 \in \mathbb{R}^m$, we get vectors $h^{ki} \in \mathbb{R}^m$ and $v^{ki} \in \mathcal{W}$ such that

$$\begin{aligned} (a_{ki}) \quad & B^T v^{ki} + A h^{ki} = B^T e^{ki}, \\ (b_{ki}) \quad & [v^{ki}]_{ki} = 0, \max_{1 \leq \ell \leq K} \|v^{ki}[\ell]\|_1 \leq aM^{-1}, \\ (c_{ki}) \quad & \max_{u \in \mathcal{U}} u^T h^{ki} + \text{Erfinv}(\epsilon) \|D^T h^{ki}\|_2 \leq a. \end{aligned} \tag{A.64}$$

2⁰. Assume now that $X_{\pm}^{ki} \neq \emptyset$. Then Y_{\pm}^{ki} are nonempty convex sets. We claim that whenever $0 < \theta < 1$, the convex compact set $\theta\mathcal{V}$ does not intersect the convex set $2Y_+^{ki}$. Indeed, if the opposite is true, there exists $v \in \mathcal{U}$ and e , $\|e\|_2 \leq \text{Erfinv}(\epsilon)$, such that $\theta(v + De) = Az$ with $z \in X_+^{ki}$. Now consider two hypotheses on the mean $\mu \in \mathbb{R}^m$ of the distribution of a Gaussian vector $\zeta \sim \mathcal{N}(\mu, DD^T)$:

$$H_+ : \mu = \theta De, \text{ and } H_- : \mu = -\theta De.$$

Let us consider the following procedure for distinguishing between these two hypotheses: given ζ , we compute $\hat{x}(\zeta)$ and accept H_+ when $[B\hat{x}(\zeta)]_{ki} > 0$, otherwise we accept H_- . We claim that this procedure rejects the true hypothesis with probability $\leq \epsilon$. Indeed, applying (4.26) to $u = -\theta v$ and $x = z$, we get

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ \|B[z - \hat{x}(Az - \theta v + D\eta)]\|_{\infty} \leq \alpha + S^{-1} L_1(Bz - [Bz]^S) \} \geq 1 - \epsilon.$$

Since $Az = \theta v + \theta De$ and $L_1(Bz - [Bz]^S) \leq \sum_{\ell \neq k} \|B^\ell z\|_{\infty} \leq M$, we get

$$\alpha + S^{-1} L_1(Bz - [Bz]^S) \leq \alpha + S^{-1} M = 2\alpha,$$

while $[Bz]_{ki} = a > 2\alpha$. It follows that if η is such that

$$\|B[z - \hat{x}(Az - \theta v + D\eta)]\|_{\infty} \leq \alpha + S^{-1} L_1(Bz - [Bz]^S),$$

then $a - [B\hat{x}(\theta De + D\eta)]_{ki} \leq 2\alpha$, whence $[B\hat{x}(\theta De + D\eta)]_{ki} > 0$. We see that

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ [B\hat{x}(\theta De + D\eta)]_{ki} > 0 \} \geq 1 - \epsilon,$$

that is, our rule for distinguishing between H_+ and H_- rejects H_+ when this hypothesis is true with probability $\leq \epsilon$. Similarly, applying (4.26) to $u = \theta v$ and $x = -z$, we get

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ \|B[-z - \hat{x}(-Az + \theta v + D\eta)]\|_{\infty} \leq \alpha + S^{-1} L_1(Bz - [Bz]^S) \} \geq 1 - \epsilon.$$

Since $-Az = -\theta v - \theta De$, we, same as above, observe that

$$\|B[-z - \hat{x}(-Az + \theta v + D\eta)]\|_{\infty} \leq \alpha + S^{-1} L_1(Bz - [Bz]^S)$$

implies that $[B\hat{x}(-\theta De + D\eta)]_{ki} < 0$, and thus

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ [B\hat{x}(-\theta De + D\eta)]_{ki} < 0 \} \geq 1 - \epsilon$$

that is, the probability to reject H_- when the hypothesis is true is $\leq \epsilon$. On the other hand, to distinguish between the hypotheses H_{\pm} via observation ζ is exactly the same as to distinguish between the distributions $\mathcal{N}(-\theta e, I_m)$ and $\mathcal{N}(\theta e, I_m)$; to do it with probabilities $\leq \epsilon$ to reject the true distribution is possible only when $\|\theta e\|_2 \geq \text{Erfinv}(\epsilon)$, which is not the case due to $\|e\|_2 \leq \text{Erfinv}(\epsilon)$ and $0 < \theta < 1$. The resulting contradiction demonstrates that $\theta\mathcal{V}$ does not intersect $2Y_+^{ki}$.

3⁰. Since $\theta\mathcal{V}$ does not intersect $2Y_+^{ki}$ when $\theta < 1$, the sets \mathcal{V} and $2Y_+^{ki}$ can be separated by a linear form, which can be normalized to be ≥ 2 on $2Y_+^{ki}$ and ≤ 2 on \mathcal{V} (recall that $0 \in \text{int } \mathcal{V}$). In other words, there exists $g = g^{ki} \in \mathbb{R}^m$ such that $\max_{v \in \mathcal{V}} g^T v \leq 2$ and $\inf_{y \in 2Y_+^{ki}} g^T y \geq 2$. Recalling the origin of \mathcal{V} , the first relation amounts to

$$\max_{u \in \mathcal{U}} u^T g + \text{Erfinv}(\epsilon) \|D^T g\|_2 \leq 1, \quad (\text{A.65})$$

while the relation $g^T y \geq 2$ for all $y \in 2Y_+^{ki} = 2AX_+^{ki}$ amounts to $f^T x \geq 1$ for all $x \in X_+^{ki}$, where $f = A^T g$. Recalling the definition of X_+^{ki} , it follows that

$$\min_x \left\{ f^T x : (e^{ki})^T Bx = a, \|B^k x\|_\infty^i + \sum_{\ell \neq k} \|B^\ell x\|_\infty \leq M \right\} \geq 1.$$

Passing to the dual problem, the latter inequality results in

$$\exists(t \in \mathbb{R}, y \in \mathcal{W}) : f = B^T y, ay_{ki} - Mt \geq 1, \sum_{j \neq i} |y_{kj}| \leq t, \max_{\ell \neq k} \|y[\ell]\|_1 \leq t. \quad (\text{A.66})$$

For the above t, y we have $0 \leq t \leq (ay_{ki} - 1)/M$, so that $y_{ki} > 0$; setting

$$h^{ki} = y_{ki}^{-1} g, \quad [v^{ki}]_{\ell j} = \begin{cases} 0, & \ell = k, j = i \\ -[y]_{\ell j} / y_{ki}, & \text{otherwise} \end{cases},$$

(A.66) combines with $f = A^T g$ to imply that $B^T e^{ki} = B^T v^{ki} + A h^{ki}$. Recall that, by construction, we have $[v^{ki}]_{ki} = 0$. Further, by (A.66) we have $\|v^{ki}[\ell]\|_1 \leq t/y_{ki} \leq a/M$, so that v^{ki}, h^{ki} satisfy (A.64.(a_{ki}), (b_{ki})). Finally, by (A.66) we have $0 < 1/y_{ki} \leq a$, which combines with (A.65) to imply (A.64.(c_{ki})).

4⁰. The bottom line here is that for every $a > 2\alpha$, $k, 1 \leq k \leq K$, and every $i, 1 \leq i \leq n_k$, there exist vectors $h^{ki} \in \mathbb{R}^m$ and $v^{ki} \in \mathcal{V}$ satisfying (A.64). It immediately follows that (A.64) can be satisfied when $a = 2\alpha$ as well. Assembling the corresponding h^{ki} and v^{ki} into the matrices

$$H = [h^{1,1}, \dots, h^{1,n_1}, h^{2,1}, \dots, h^{2,n_2}, \dots, h^{K,1}, \dots, h^{K,n_K}], \\ V = [(v^{1,1})^T; \dots; (v^{1,n_1})^T; (v^{2,1})^T; \dots; (v^{2,n_2})^T; \dots; (v^{K,1})^T; \dots; (v^{K,n_K})^T],$$

we get (4.27) as an immediate consequence of (A.64) with a set to 2α . ■

A.6 Proofs for section 5.1

Proof of Proposition 5.1 Let $V^\ell = [V^{1\ell}; \dots; V^{K\ell}]$, $1 \leq \ell \leq K$, be the “stripes” of V . Given $x \in \mathbb{R}^n$, setting $w = Bx$, and using the relation (5.28), we have

$$w = Bx = [VB + H^T A]x = Vw + H^T Ax,$$

whence,

$$\begin{aligned} L_{s,q}(w) &= L_{s,q}(Vw + H^T Ax) \leq L_{s,q}(Vw) + L_{s,q}(H^T Ax) \\ &= L_{s,q} \left(\sum_{\ell=1}^K V^\ell w[\ell] \right) + s^{\frac{1}{q}} L_\infty(H^T Ax) \\ &\leq \sum_{\ell=1}^K \| [\|V^{1\ell} w[\ell]\|_{(1)}; \dots; \|V^{K\ell} w[\ell]\|_{(K)}] \|_{s,q} + s^{\frac{1}{q}} L_\infty(H^T Ax) \\ &\leq \sum_{\ell=1}^K \nu_{s,q}^*(V) \|w[\ell]\|_{(\ell)} + s^{\frac{1}{q}} L_\infty(H^T Ax) = \nu_{s,q}^*(V) L_1(w) + s^{\frac{1}{q}} L_\infty(H^T Ax) \end{aligned}$$

proving (5.29). ■

Proof of Proposition 5.2 To verify (5.32), note that for every k and every ℓ , we have

$$0 \leq \|V^{k\ell}w[\ell]\|_{(k)} \leq \|w[\ell]\|_{(\ell)} \max_{w \in \mathbb{R}^{n_\ell}: \|w\|_{(\ell)} \leq 1} \|V^{k\ell}w\|_{(k)} = \|w[\ell]\|_{(\ell)} \|V^{k\ell}\|_{(\ell,k)} = \|w[\ell]\|_{(\ell)} \Omega_{k\ell}.$$

Since for any two nonnegative vectors, a, b satisfying $a_i \leq b_i \forall i$, we have $\|a\|_{s,q} \leq \|b\|_{s,q}$, we get

$$\|[\|V^{1\ell}w[\ell]\|_{(1)}; \dots; \|V^{K\ell}w[\ell]\|_{(K)}]\|_{s,q} \leq \|w[\ell]\|_{(\ell)} \|\text{Col}_\ell[\Omega]\|_{s,q}.$$

By taking the maximum of both sides first with respect to $w[\ell]$ subject to the constraint that $\|w[\ell]\|_{(\ell)} \leq 1$, and then over $1 \leq \ell \leq K$, we arrive at (5.32) implying that $\nu_{s,q}^*(V) \leq \hat{\nu}_{s,q}(V)$. ■

A.7 Derivations for section 5.2

Derivation of $\mathbf{S}_{\kappa,q}$ in all three cases (i.e., $r = \infty$, $r = 2$, $r = 1$) is quite straightforward — we just plug into (5.33) the description of $\hat{\nu}_{s,q}(\cdot)$. Let us focus on the derivation of \mathbf{R}_ϵ .

The case of $r = \infty$. In this case, $L_\infty(H^T(u + \xi)) = \|H^T(u + D\eta)\|_\infty$, so that the requirement

$$L_\infty(H^T(u + D\eta)) \leq \rho \quad \forall u \in \mathcal{U} \tag{A.67}$$

amounts to

$$\max_{\substack{1 \leq t \leq n_k \\ 1 \leq k \leq K}} \left\{ \|E^T h^{kt}\|_2 + |\eta^T D^T h^{kt}| \right\} \leq \rho. \tag{A.68}$$

The condition (5.35. \mathbf{R}_ϵ) is a natural sufficient condition for (A.68) to be satisfied with probability $\geq 1 - \epsilon$ when $\eta \sim \mathcal{N}(0, I_m)$ (cf. the derivations in section 4.1).

The case of $r = 2$. Here a slightly conservative (tight within factor 2) sufficient condition for (A.67) reads

$$\max_{1 \leq k \leq K} \left\{ \|E^T H[k]\|_2 + \|H^T[k] D \eta\|_2 \right\} \leq \rho. \tag{A.69}$$

A simple sufficient condition for (A.69) to be satisfied with probability $\geq 1 - \epsilon$ when $\eta \sim \mathcal{N}(0, I_m)$ is for the probability of the event

$$\left\{ \eta : \|E^T H[k]\|_2 + \sqrt{\eta^T D^T H[k] H^T[k] D \eta} \leq \rho \right\} \tag{A.70}$$

to be $\geq 1 - \epsilon/K$ for every $k \leq K$. For a given k , invoking the Schur Complement Lemma, this condition is the same as the existence of a symmetric matrix W_k and a real α_k such that

$$\begin{bmatrix} W_k & D^T H[k] \\ H^T[k] D & \alpha_k I_{n_k} \end{bmatrix} \succeq 0 \text{ \& } \|E^T H[k]\|_2 + \alpha_k \leq \rho \text{ \& } \text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ \eta^T W_k \eta \leq \alpha_k \} \geq 1 - \epsilon/K. \tag{A.71}$$

By [5, Proposition 4.5.10], for every symmetric $m \times m$ matrix W , setting $\lambda = \lambda(W)$, one has

$$\forall \varkappa \geq 0 : \text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ \eta^T W \eta > \text{Tr}(W) + \varkappa \|\lambda\|_2 \} \leq \exp \left\{ -\frac{\varkappa^2 \|\lambda\|_2}{4(2\|\lambda\|_2 + \|\lambda\|_\infty \varkappa)} \right\}.$$

The smallest $\varkappa \geq 0$ for which the right hand side of this inequality is $\leq \epsilon/K$ satisfies

$$\varkappa \|\lambda\|_2 = 2\delta \|\lambda\|_\infty + \sqrt{4\|\lambda\|_\infty^2 + 8\|\lambda\|_2^2 \delta}, \quad \delta = \ln \frac{K}{\epsilon},$$

and we conclude that the condition

$$\text{Tr}(W_k) + 2\delta\|\lambda(W_k)\|_\infty + \sqrt{4\|\lambda(W_k)\|_\infty^2 + 8\delta\|\lambda(W_k)\|_2^2} \leq \alpha_k$$

is sufficient for the validity of the third condition in (A.71). This observation implies straightforwardly that the condition (5.36.R_ε) indeed is sufficient for the validity of (A.69) with probability $\geq 1 - \epsilon$, which is all we need to prove.

The case of $r = 1$. Here a slightly conservative (tight within factor 2) sufficient condition for (A.67) reads

$$\max_{1 \leq k \leq K} \left\{ \max_{\|v\|_2 \leq 1} \|H^T[k]Ev\|_1 + \|H^T[k]D\eta\|_1 \right\} \leq \rho. \quad (\text{A.72})$$

A natural way to enforce the validity of this condition with probability $\geq 1 - \epsilon$ when $\eta \sim \mathcal{N}(0, I_m)$ is to ensure that for every $k \leq K$ we have

$$\text{Prob}_{\eta \sim \mathcal{N}(0, I_m)} \{ \eta : \|H^T[k]E\|_{2,1} + \|H^T[k]D\eta\|_1 \leq \rho \} \geq 1 - \epsilon/K. \quad (\text{A.73})$$

A natural upper bound on the $1 - \epsilon/K$ -quantile of $\|H^T[k]D\eta\|_1$ is $\text{Erfinv}\left(\frac{\epsilon}{2Kn_k}\right) \sum_{t=1}^{n_k} \|D^T h^{kt}\|_2$, so that the relation

$$\|H^T[k]E\|_{2,1} + \text{Erfinv}\left(\frac{\epsilon}{2Kn_k}\right) \leq \rho. \quad (\text{A.74})$$

The latter constraint, while convex in variables H, ρ , is intractable, since the norm $\|H^T[k]E\|_{2,1}$ is difficult to compute. This norm, however, admits a tight within the factor $\sqrt{\pi/2} \approx 1.29$ efficiently computable upper bound on $\|\cdot\|_{2,1}$, due to Yu. Nesterov (see [32, Theorem 13.2.4]),⁸ namely, for every $G \in \mathbb{R}^{p \times q}$ one has

$$\|G\|_{2,1} \leq \min_{\lambda \in \mathbb{R}^p, \mu \in \mathbb{R}} \left\{ \frac{1}{2} \sum_i \lambda_i + \frac{1}{2} \mu : \begin{bmatrix} \text{Diag}\{\lambda\} & G \\ G^T & \mu I_q \end{bmatrix} \succeq 0 \right\} \quad (\text{A.75})$$

Replacing in (A.74) the quantity $\|H^T[K]E\|_{2,1}$ with its upper bound, we arrive at (5.37).

A.8 Proof of Proposition 5.3

Let H, V satisfy (5.38) with $B = I$, that is, $V = I_n - H^T A$, V is $n \times n$, and H is $m \times n$. Observe, first, that $sd \leq m$; indeed, otherwise, by dimension argument, we could find s -block-sparse signal $x \neq 0$ such that $Ax = 0$, meaning that there is no way to recover x even from the exact observation of Ax and thus our sufficient condition for the validity of ℓ_1 -recovery cannot hold true. Now let us set $\bar{K} = \text{Ceil}(2m/d)$ and $\bar{n} = \bar{K}d$. Note that $\bar{K} \leq K$ due to $n = Kd \geq 2m$, whence $\bar{n} \leq n$ and of course $\bar{n} < 2m + d$ due to $(\bar{K} - 1)d < 2m$. Now let $\bar{V} = [V^{k\ell}]_{k,\ell=1}^{\bar{K}}$ be the $\bar{n} \times \bar{n}$ “North-Western” block of V , \bar{A} be $m \times \bar{n}$ matrix comprised of the first \bar{n} columns of A , and \bar{H} be $m \times \bar{n}$ matrix comprised of the first \bar{n} columns of H . Since $V = I_n - H^T A$, we have $\bar{V} = I_{\bar{n}} - \bar{H}^T \bar{A}$. Moreover, (5.38) implies that

$$\underbrace{\|V^{1\ell}\|_{r,r}; \|V^{2\ell}\|_{r,r}; \dots; \|V^{\bar{K}\ell}\|_{r,r}}_{f^\ell \in \mathbb{R}^{\bar{K}}} \|_{s,q} \leq \alpha := \frac{1}{2} s^{\frac{1}{q}-1}, \quad 1 \leq \ell \leq \bar{K}. \quad (\text{A.76})$$

Now, $\bar{V} = I_{\bar{n}} - \bar{H}^T \bar{A}$, and since the rank of $H^T A$ is $\leq m$, at least $\bar{n} - m$ singular values of \bar{V} are ≥ 1 , and therefore the squared Frobenius norm $\|\bar{V}\|_F^2$ of \bar{V} is at least $\bar{n} - m$. On the other hand, we can upper-bound this squared norm as follows. It is immediately seen that with the values of r in question, for every

⁸To see why the right-hand side of (A.75) is indeed an upper bound for $\|G\|_{2,1}$, take $v, \|v\|_2 \leq 1$, set $g = Gv$ and note that by the Schur Complement Lemma for λ, μ feasible for the right hand side problem in (A.75) we have $\sum_i g_i^2 / \lambda_i \leq \mu v^T v \leq \mu$; noting that $[\sum_i |g_i|]^2 \leq [\sum_i \lambda_i] [\sum_i g_i^2 / \lambda_i]$ whenever $\lambda_i \geq 0$, we conclude that $\|g\|_1 \leq \sqrt{\mu \sum_i \lambda_i} \leq \frac{1}{2} \mu + \frac{1}{2} \sum_i \lambda_i$. Recalling the origin of g , we conclude that the right hand side in (A.75) indeed is an upper bound on $\|G\|_{2,1}$.

$d \times d$ block $V^{k\ell}$ in \bar{V} we have $\|V^{k\ell}\|_F \leq \sqrt{d}\|V^{k\ell}\|_{r,r}$. Since for every \bar{K} -dimensional vector f one has $\|f\|_2^2 \leq \max\left[\frac{\bar{K}}{s^{2/q}}, 1\right] \|f\|_{s,q}^2$, we now get

$$\begin{aligned} \bar{n} - m &\leq \|\bar{V}\|_F^2 = \sum_{\ell=1}^{\bar{K}} \sum_{k=1}^{\bar{K}} \|V^{k\ell}\|_F^2 \leq d \sum_{\ell=1}^{\bar{K}} \sum_{k=1}^{\bar{K}} \|V^{k\ell}\|_{r,r}^2 = d \sum_{\ell=1}^{\bar{K}} \|f^\ell\|_2^2 \\ &\leq d \sum_{\ell=1}^{\bar{K}} \max\left[\frac{\bar{K}}{s^{2/q}}, 1\right] \|f^\ell\|_{s,q}^2 \leq \max\left[\frac{\bar{K}}{s^{2/q}}, 1\right] d\bar{K}\alpha^2 = \frac{1}{4}\bar{n} \max\left[d^{-1}\bar{n}s^{-2}, s^{\frac{2}{q}-2}\right]. \end{aligned}$$

We conclude that either

$$s^{\frac{2}{q}-2} \geq \frac{4(\bar{n} - m)}{\bar{n}}, \quad (a)$$

or

$$s^2 \leq \frac{\bar{n}^2}{4d(\bar{n} - m)}. \quad (b)$$

Since $\bar{n} \geq 2m$, the right hand side in (a) is ≥ 2 ; thus, (a) is impossible. Taking into account that, as we have seen, $\bar{n} - m \geq m$ and $\bar{n} < 2m + d \leq 3m$, (b) implies (5.39). \blacksquare

A.9 Proof of Proposition 5.4

1⁰. The proof of (i) is immediate. Indeed, let us look at the blocks $V^{k\ell}$ in the matrix $V = I - H^T A$. When $k = \ell$, we have $V^{kk} = (1 - \frac{1}{\mu+1})I_{n_k} = \frac{\mu}{1+\mu}I_{n_k}$ by definition of C_k , whence $\|V^{kk}\|_{2,2} = \frac{\mu}{1+\mu}$. When $k \neq \ell$, we have $V^{k\ell} = -\frac{1}{\mu+1}C_k^{-1}A^T[k]A[\ell]$, that is, $\|V^{k\ell}\|_{2,2} \leq \frac{\mu}{1+\mu}$ by the definition of μ . The bottom line is that $\|V^{k\ell}\|_{2,2} \leq \frac{\mu}{1+\mu}$, and thus, in the notation of Proposition 5.1, $\hat{\nu}_{s,\infty}(V) \leq \frac{\mu}{1+\mu}$. We see that (5.44) indeed implies the validity of (5.33), which is all we need.

2⁰. Let us justify the bound on the mutual block-incoherence stated in (ii). Specifically, let $n_k = d$, $1 \leq k \leq K$, let A be a matrix with unit columns, let ν, μ_B be given by (5.41), and let the first relation in (5.42) be satisfied; we want to prove that the mutual block-incoherence of A is $\leq \mu = \frac{d\mu_B}{1-(d-1)\nu}$. This is immediate: since the columns of A have unit ℓ_2 -norms, the diagonal entries in the $d \times d$ symmetric matrix $C_k = A^T[k]A[k]$ are equal to 1, while the moduli of the off-diagonal entries in C_k do not exceed ν by the definition of ν . It immediately follows that the minimal eigenvalue of C_k is $\geq 1 - (d-1)\nu$, and thus C_k is positive definite (by the first relation in (5.42)) and $\sigma_{\max}(C_k^{-1}) \leq [1 - (d-1)\nu]^{-1}$. Therefore, whenever $k \neq \ell$, we have $\sigma_{\max}(C_k^{-1}A^T[k]A[\ell]) \leq \sigma_{\max}(C_k^{-1})\sigma_{\max}(A^T[k]A[\ell]) \leq (1 - (d-1)\nu)^{-1}\sigma_{\max}(A^T[k]A[\ell]) \leq \frac{d\mu_B}{1-(d-1)\nu}$, where the concluding inequality is a consequence of the definition of μ_B , see (5.41). Recalling the definition of the mutual block-incoherence, we arrive at the desired conclusion. \blacksquare

A.10 Proof of Proposition 5.5

1⁰ Let $\xi \sim \mathcal{N}(0, m^{-1}I_m)$ and $\eta \sim \mathcal{N}(0, m^{-1}I_m)$ be two independent normal vectors. Then for any $0 < \lambda \leq \sqrt{m}$ we have

$$\text{Prob}\left\{|\xi^T \eta| \geq \frac{\lambda}{\sqrt{m}}\right\} \leq 2e^{-\frac{\lambda^2}{3}}; \quad \text{Prob}\left\{|\xi^T \xi - 1| \geq \frac{\lambda}{\sqrt{m}}\right\} \leq 2e^{-\frac{\lambda^2}{8}}. \quad (\text{A.77})$$

Indeed, by the standard argument, we have for $0 \leq t < \sqrt{m}$: $\ln \left(E \left[e^{t\sqrt{m}\xi^T\eta} \right] \right) = -\frac{m}{2} \ln(1 - \frac{t^2}{m})$. We conclude that when $t^2 \leq \frac{m}{2}$, $\ln \left(E \left[e^{t\sqrt{m}\xi^T\eta} \right] \right) \leq \frac{3}{4}t^2$, and for $\lambda^2 \leq \frac{9m}{8}$

$$\text{Prob}\{|\xi^T\eta| > \frac{\lambda}{\sqrt{m}}\} = 2\text{Prob}\{\xi^T\eta > \frac{\lambda}{\sqrt{m}}\} \leq 2 \min_{|t| \leq \sqrt{m}/2} e^{\frac{3}{4}t^2 - \lambda t} = 2e^{-\frac{\lambda^2}{3}}.$$

Further, we have for $t \leq \frac{\sqrt{m}}{4}$:

$$\begin{aligned} \ln \left(E \left[e^{t\sqrt{m}(\xi^T\xi - 1)} \right] \right) &= -\frac{m}{2} \ln(1 - \frac{2t}{\sqrt{m}}) - t\sqrt{m} \leq 2t^2, \\ \ln \left(E \left[e^{t\sqrt{m}(1 - \xi^T\xi)} \right] \right) &= -\frac{m}{2} \ln(1 + \frac{2t}{\sqrt{m}}) + t\sqrt{m} \leq t^2. \end{aligned}$$

Consequently, when $\lambda \leq \sqrt{m}$,

$$\text{Prob}\{|\xi^T\xi - 1| > \frac{\lambda}{\sqrt{m}}\} \leq 2 \min_{|t| \leq \sqrt{m}/4} e^{2t^2 - \lambda t} = 2e^{-\frac{\lambda^2}{8}},$$

and we arrive at (A.77).

2^o Now, let $S = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$, and let \mathcal{D}_ϵ be a minimal ϵ -net, w.r.t. $\|\cdot\|_2$, in S , and let \mathcal{N}_ϵ be the cardinality of \mathcal{D}_ϵ . Further, let $1 \leq \ell, k \leq K$ and let $A[\ell]$ and $A[k]$ be the submatrices of A comprised of columns of A with indices in the ℓ -th and the k -th blocks. Finally, let $G^\ell = A^T[\ell]A[\ell]$, $C^\ell = A[\ell]^T A[\ell] - I_d$, and let $x, y \in S$ be deterministic. When invoking (A.77) for $\xi = A[\ell]x$ and $\eta = A[k]y$ we conclude that

$$\text{Prob} \left\{ |x^T A[\ell]^T A[k]y| \geq \frac{\lambda}{\sqrt{m}} \right\} \leq 2e^{-\frac{\lambda^2}{3}}, \quad k \neq \ell; \quad \text{Prob} \left\{ |x^T C^\ell x| \geq \frac{\lambda}{\sqrt{m}} \right\} \leq 2e^{-\frac{\lambda^2}{8}}. \quad (\text{A.78})$$

We claim that

$$\left\{ |v^T C^\ell v| \leq \Delta \quad \forall v \in \mathcal{D}_\epsilon \right\} \Rightarrow \left\{ \sigma_{\max}(C^\ell) \leq (1 - 2\epsilon)^{-1} \Delta \right\} \quad (\text{A.79})$$

where $\sigma_{\max}(\cdot)$ is the spectral norm. Indeed, let the premise in (A.79) hold true. Recall that C^ℓ is symmetric, so let $\bar{u} \in S$ be such that $|\bar{u}^T C^\ell \bar{u}| = \sigma_{\max}(C^\ell)$. There exists $v \in \mathcal{D}_\epsilon$ such that $\|\bar{u} - v\|_2 \leq \epsilon$, whence $\sigma_{\max}(C^\ell) = |\bar{u}^T C^\ell \bar{u}| \leq 2\sigma_{\max}(C^\ell)\|\bar{u} - v\|_2 + |v^T C^\ell v| \leq 2\sigma_{\max}(C^\ell)\epsilon + \Delta$ (since the quadratic form $z^T C^\ell z$ is Lipschitz continuous, with constant $2\sigma_{\max}(C^\ell)$ w.r.t. $\|\cdot\|_2$, on S), whence $\sigma_{\max}(C^\ell) \leq (1 - 2\epsilon)^{-1} \Delta$. By similar reasons when $k \neq \ell$ we have

$$\left\{ |v^T A[\ell]^T A[k]v| \leq \Delta \quad \forall v, v' \in \mathcal{D}_\epsilon \right\} \Rightarrow \left\{ \sigma_{\max}(A[\ell]^T A[k]) \leq (1 - 2\epsilon)^{-1} \Delta \right\} \quad (\text{A.80})$$

Indeed, let $u, u' \in S$ be such that $u^T A[\ell]^T A[k]u' = \sigma_{\max}(A[\ell]^T A[k])$, and if $v, v' \in \mathcal{D}_\epsilon$ satisfy $\|u - v\|_2 \leq \epsilon$, $\|u' - v'\|_2 \leq \epsilon$, then

$$u^T A[\ell]^T A[k]u' \leq v^T A[\ell]^T A[k]v' + \sigma_{\max}(A[\ell]^T A[k])(\|u - v\|_2 + \|u' - v'\|_2) \leq \Delta + 2\sigma_{\max}(A[\ell]^T A[k])\epsilon.$$

3⁰ We can straightforwardly build an ϵ -net \mathcal{D}' in S in such a way that the $\|\cdot\|_2$ -distance between every two distinct points of the net is $> \epsilon$, so that the balls $B_v = \{z \in \mathbb{R}^s : \|z - v\|_2 \leq \epsilon/2\}$ with $v \in \mathcal{D}'$ are mutually disjoint. Since the union of these balls belongs to $B = \{z \in \mathbb{R}^s : \|z\|_2 \leq 1 + \epsilon/2\}$, we get $\text{Card}(\mathcal{D}')(\epsilon/2)^d \leq (1 + \epsilon/2)^d$, that is, $\mathcal{N}_\epsilon \leq \text{Card}(\mathcal{D}') \leq (1 + 2/\epsilon)^d$.

Let us set $\epsilon = \frac{1}{4}$. When using the second inequality of (A.78), we conclude that the probability of violating the premise in (A.79) with $\Delta = 2\lambda m^{-1/2}$ and $\lambda \leq \sqrt{m}$ does not exceed $2 \exp\{-\frac{\lambda^2}{8} + d \ln[1 + 2\epsilon^{-1}]\} = 2 \exp\{-\frac{\lambda^2}{8} + d \ln[9]\}$. Setting $\lambda = \sqrt{m}/4$, we conclude that when $m \geq 128 \ln(4n^2/d) + 256d \ln(3)$, one has

$$\text{Prob} \left\{ \max_{1 \leq \ell \leq K} \sigma_{\max}(C^\ell - I_d) > \frac{1}{2} \right\} \leq \frac{1}{2n},$$

that is, with probability at least $1 - \frac{1}{2n}$ all the matrices $C^\ell + I_d = A^T[\ell]A[\ell] =: G^\ell$ are invertible with $\max_{1 \leq \ell \leq K} \sigma_{\max}(G^\ell)^{-1} \leq 2$.

In the same way, using (A.80) and the first inequality of (A.78) we obtain

$$\lambda := \sqrt{3 \ln(4n^3/d^2) + 6d \ln(3)} \leq \sqrt{m} \Rightarrow \text{Prob} \left\{ \sigma_{\max}([A^T[k]A[\ell]]) \leq \frac{2\lambda}{\sqrt{m}} \quad \forall k \neq \ell \right\} \geq 1 - \frac{1}{2n}.$$

We conclude that for properly chosen absolute constants C_1, C_2 under the condition $m \geq C_1 \max[d, \ln(n)]$ we have

$$\text{Prob} \left\{ \sigma_{\max}(A^T[k]A[\ell]) \leq C_2 \sqrt{[d + \ln(n)]/m} \quad \forall k \neq \ell, \quad \sigma_{\max}((A^T[\ell]A[\ell])^{-1}) \leq 2 \quad \forall \ell \right\} \geq 1 - \frac{1}{n},$$

so that by the definition of the mutual block-incoherence

$$\text{Prob}\{A : \mu(A) \leq C_2 \sqrt{[d + \ln(n)]/m}\} \geq 1 - \frac{1}{n}. \quad \blacksquare$$

A.11 Proof of Lemma 6.1

The proof below follows the lines of the proof of Proposition 10 of [22]. Let Ξ^+ be given by (6.46.c). Let us fix $\xi \in \Xi^+$, $u \in \mathcal{U}$ and $x \in \mathbb{R}^n$ such that $L_1(Bx - [Bx]^s) \leq v$, and let $\eta = u + \xi$, so that $L_\infty(H^T \eta) \leq \rho$ due to $\xi \in \Xi^+$. We will proceed by induction. First, let us show that (a_{k-1}, b_{k-1}) implies (a_k, b_k) . Thus, assume that (a_{k-1}, b_{k-1}) holds true. Let $z^{(k-1)} = x - v^{(k-1)}$. With g and Δ as defined at k -th step of the algorithm, we have

$$\begin{aligned} Bz^{(k-1)} - g &= Bx - Bv^{(k-1)} - H^T(y - Av^{(k-1)}) = (B - H^T A)(x - v^{(k-1)}) - H^T \eta \\ &= VBz^{(k-1)} - H^T \eta, \end{aligned}$$

by (6.46.a). Now, due to (6.46.b) and the fact that $L_\infty(H^T \eta) \leq \rho$ by (6.46.c), we have

$$L_\infty(Bz^{(k-1)} - g) \leq L_\infty(VBz^{(k-1)}) + L_\infty(H^T \eta) \leq \bar{\gamma} L_1(Bz^{(k-1)}) + \rho \leq \bar{\gamma} \alpha_{k-1} + \rho := \nu, \quad (\text{A.81})$$

where the last inequality is due to (b_{k-1}) . Let us fix $j \in \{1, \dots, K\}$; let us consider the case $\|\cdot\|_{(j)} = \|\cdot\|_2$. Observe that $\Delta[j]$ is the closest to 0 point of the Euclidean ball $S_j = \{w \in \mathbb{R}^{n_j} : \|w - g[j]\|_2 \leq \nu\}$. By the properties of the Euclidean projection, for any $w \in S_j$, $\|w - \Delta[j]\|_2 \leq \|w\|_2$, and, since by (A.81) $(Bz^{(k-1)})[j] \in S_j$, we have

$$\|(Bz^{(k)})[j]\|_2 = \|(Bz^{(k-1)} - \Delta)[j]\|_2 \leq \|(Bz^{(k-1)})[j]\|_2 \leq \|(Bx)[j]\|_2, \quad (\text{A.82})$$

since by (a_{k-1}) , $\|(Bz^{(k-1)})[j]\|_2 \leq \|(Bx)[j]\|_2$. In the case where $\|\cdot\|_{(j)} = \|\cdot\|_\infty$ we note that Δ_{ji} is the closest to 0 point of the interval $S_{ji} = [g_{ji} - \nu, g_{ji} + \nu]$, $i = 1, \dots, n_j$, and by (A.81), the segments S_{ji} cover $(Bz^{(k-1)})_{ji}$. We conclude that $\Delta_{ji} \in \text{Conv}\{0, (Bz^{(k-1)})_{ji}\}$, so that

$$\|(Bz^{(k)})[j]\|_\infty = \|(Bz^{(k-1)} - \Delta)[j]\|_\infty \leq \|(Bz^{(k-1)})[j]\|_\infty \leq \|(Bx)[j]\|_\infty$$

where the last inequality is by (a_{k-1}) . Together with (A.82) this implies (a_k) . Further, we have by (A.81):

$$L_\infty(Bz^{(k)}) = L_\infty(Bz^{(k-1)} - \Delta) \leq 2\nu = 2\bar{\gamma}\alpha_{k-1} + 2\rho.$$

Let now $I \subset \{1, \dots, K\}$ be the set of indexes of the s largest in magnitude blocks in Bx , and $J = \{1, \dots, K\} \setminus I$. Now by (a_k) ,

$$\begin{aligned} L_1(Bz^{(k)}) &= L_1([Bz^{(k)}]_I) + L_1([Bz^{(k)}]_J) \leq \sum_{j \in I} \|(Bz^{(k)})[j]\|_{(j)} + \sum_{j \notin I} \|(Bx)[j]\|_\infty \\ &\leq sL_\infty(Bz^{(k-1)} - \Delta) + v \leq 2s\nu + v = \alpha_k. \end{aligned}$$

we conclude that (b_k) holds, what completes the induction step.

It remains to show that (a_0, b_0) holds true. Since (a_0) is evident, all we need is to justify (b_0) . Let

$$\alpha_* = L_1(Bx),$$

and let $g = H^T y$. Same as above (cf. (A.81)), we have for all i :

$$L_\infty(Bx - g) \leq \bar{\gamma}\alpha_* + \rho.$$

Then

$$\begin{aligned} \alpha_* &= L_1(Bx) = L_1([Bx]_I) + L_1(Bx - [Bx]^s) \\ &\leq \sum_{j \in I} [\|g[j]\|_{(j)} + \bar{\gamma}\alpha_* + \rho] + v \leq L_{s,1}(g) + s\bar{\gamma}\alpha_* + s\rho + v. \end{aligned}$$

Hence

$$\alpha_* \leq \alpha_0 = \frac{L_{s,1}(g) + s\rho + v}{1 - s\bar{\gamma}},$$

what implies (b_0) . ■